DOCUMENT RESUME

ED 327 567 TM 015 998

TITLE	Oversight Hearing on Testing/Assessment/Evaluation To Improve Learning in Our Schools. Hearing before the Subcommittee on Elementary, Secondary, and Vocational Education of the Committee on Education and Labor, House of Representatives. One Hundred First Congress, Second Session.
INSTITUTION	Congress of the U.S., Washington, D.C. House Committee on Education and Labor.
PUB DATE	7 Jun 90
NOTE	130p.; Serial No. 101-111.
AVAILABLE FROM	Superintendent of Documents, Congressional Sales Office, U.S. Government Printing Office, Washington, DC 20402.
PUB TYPE	Legal/Legislative/Regulatory Materials (090) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Educational Assessment; Educational Improvement; Elementary Secondary Education; *Evaluation Methods; Federal Government; National Surveys; *Standardized Tests; Student Evaluation; Test Construction; Testing Problems; Test Results; *Test Use; Vocational Education

ABSTRACT

This document provides statements presented at the oversight hearing on testing and assessment evaluation to improve learning in the nation's schools. Walter Haney summarizes the National Commission on Testing and Public Policy's recent report, which concluded that while well-designed and responsibly used assessment is an important source of information, there is too much national reliance on imperfect and unfair measures. Burton W. Faldet, on behalf of the Association of American Publishers, discusses test use and the importance of improving test quality. Walter E. Faithorn, Jr., representing Friends of Education, highlights the group's misgivings about test use and the misuse of standardized testing. Ramsay Selden, of the State Education Assessment Center of the Council of Chief State School Officers, indicates that better testing practices are within reach. Much of the discussion subsequent to these statements centers on the connections among test publishers, textbook publishers, and the construction of tests. Additional prepared statements, letters, and supplemental materials are included from the following persons and agencies: (1) the American Federation of Teachers; (2) the American Psychological Association; (3) the Council for Basic Education; (4) Christopher T. Cross, Office of Educational Research and Improvement; (5) Frederick H. Dietrich, the College Board; (6) Emerson J. Elliot, National Center for Education Statistics; (7) Matthew G. Martinez, Representative from California; (8) Monty Neill, National Center for Fair and Open Testing; and (9) Daniele G. Rodamar, American University. (SLD)

TM

OVERSIGHT HEARING ON TESTING/ASSESSMENT/ EVALUATION TO IMPROVE LEARNING IN OUR SCHOOLS

ED327567

HEARING

BEFORE THE

SUBCOMMITTEE ON ELEMENTARY, SECONDARY, AND VOCATIONAL EDUCATION

OF THE

COMMITTEE ON EDUCATION AND LABOR HOUSE OF REPRESENTATIVES

ONE HUNDRED FIRST CONGRESS

SECOND SESSION

HEARING HELD IN WASHINGTON, DC, JUNE 7, 1990

Serial No. 101-111

Printed for the use of the Committee on Education and Labor

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

U.S. GOVERNMENT PRINTING OFFICE
34-661 WASHINGTON : 1990

For sale by the Superintendent of Documents, Congressional Sales Office
U.S. Government Printing Office, Washington, DC 20402

TM015998

COMMITTEE ON EDUCATION AND LABOR

AUGUSTUS F. HAWKINS, California, *Chairman*

WILLIAM D. FORD, Michigan
JOSEPH M. GAYDOS, Pennsylvania
WILLIAM (BILL) CLAY, Missouri
GEORGE MILLER, California
AUSTIN J. MURPHY, Pennsylvania
DALE E. KILDEE, Michigan
PAT WILLIAMS, Montana
MATTHEW G. MARTINEZ, California
MAJOR R. OWENS, New York
CHARLES A. HAYES, Illinois
CARL C. PERKINS, Kentucky
THOMAS C. SAWYER, Ohio
DONALD M. PAYNE, New Jersey
NITA M. LOWEY, New York
GLENN POSHARD, Illinois
JOLENE UNSOELD, Washington
CRAIG A. WASHINGTON, Texas
JOSÉ E. SERRANO, New York
JAIME B. FUSTER, Puerto Rico
JIM JONTZ, Indiana
KWEISI MFUME, Maryland

WILLIAM F. GOODLING, Pennsylvania
E. THOMAS COLEMAN, Missouri
THOMAS E. PETRI, Wisconsin
MARGE ROUKEMA, New Jersey
STEVE GUNDERSON, Wisconsin
STEVE BARTLETT, Texas
THOMAS J. TAUKE, Iowa
HARRIS W. FAWELL, Illinois
PAUL B. HENRY, Michigan
FRED GRANDY, Iowa
CASS BALLENGER, North Carolina
PETER SMITH, Vermont
TOMMY F. ROBINSON, Arkansas

SUBCOMMITTEE ON ELEMENTARY, SECONDARY, AND VOCATIONAL EDUCATION

AUGUSTUS F. HAWKINS, California, *Chairman*

WILLIAM D. FORD, Michigan
GEORGE MILLER, California
DALE E. KILDEE, Michigan
PAT WILLIAMS, Montana
MATTHEW G. MARTINEZ, California
CARL C. PERKINS, Kentucky
CHARLES A. HAYES, Illinois
THOMAS C. SAWYER, Ohio
MAJOR R. OWENS, New York
DONALD M. PAYNE, New Jersey
NITA M. LOWEY, New York
GLENN POSHARD, Illinois
JOLENE UNSOELD, Washington
JOSÉ E. SERRANO, New York

WILLIAM F. GOODLING, Pennsylvania
HARRIS W. FAWELL, Illinois
FRED GRANDY, Iowa
PETER SMITH, Vermont
STEVE BARTLETT, Texas
STEVE GUNDERSON, Wisconsin
THOMAS E. PETRI, Wisconsin
MARGE ROUKEMA, New Jersey
E. THOMAS COLEMAN, Missouri

CONTENTS

	Page
Hearing held in Washington, DC, June 7, 1990	1
Statement of:	
Haney, Dr. Walter, Boston College; Dr. Burton W. Faldet, Test Consultants, Ltd.; Walter E. Faithorn, Jr., Business Executive and Volunteer Teacher at the University of the District of Columbia; and Ramsay Selden, State Education Assessment Center Council of Chief State School Officers	5
Prepared statements, letters, supplemental materials, et cetera:	
American Federation of Teachers, AFL-CIO, prepared statement of	107
American Psychological Association, prepared statement of	101
Anrig, Gregory R., President, Educational Testing Service, letter dated July 19, 1990, to Hon. Augustus F. Hawkins, enclosing material for the record	86
Council for Basic Education, prepared statement of	104
Cross, Christopher T., Assistant Secretary for Educational Research and Improvement, U.S. Department of Education, letter dated July 9, 1990, to Hon. Augustus F. Hawkins, enclosing material for the record	79
Dietrich, Frederick H., Vice President, Guidance, Access, and Assessment Services, The College Board, prepared statement of	123
Elliot, Emerson J., Acting Commissioner, National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education, letter dated July 12, 1990, to Hon. Augustus F. Hawkins, enclosing responses for the record	81
Faithorn, Walter E., Jr., Business Executive and Volunteer Teacher at the University of the District of Columbia, prepared statement of	34
Faldet, Dr. Burton W., Test Consultants, Ltd., prepared statement with attachments	11
Friends for Education, Inc., prepared statement of	73
Martinez, Hon. Matthew G., a Representative in Congress from the State of California, prepared statement of	2
Neill, Monty, Ed.D., Associate Director, National Center for Fair and Open Testing, letter dated June 21, 1990, to Hon. Augustus F. Hawkins, enclosing material for the record	93
Rodamar, Daniele Ghiolfi, Assistant Professor, American University, prepared statement of	110
Selden, Ramsay, State Education Assessment Center, Council of Chief State School Officers, prepared statement of	49

OVERSIGHT HEARING ON TESTING/ASSESSMENT/EVALUATION TO IMPROVE LEARNING IN OUR SCHOOLS

THURSDAY, JUNE 7, 1990

HOUSE OF REPRESENTATIVES,
SUBCOMMITTEE ON ELEMENTARY, SECONDARY,
AND VOCATIONAL EDUCATION,
COMMITTEE ON EDUCATION AND LABOR,
Washington, DC.

The subcommittee met, pursuant to notice, at 9:50 a.m., in Room 2175, Rayburn House Office Building, Hon. Augustus F. Hawkins [Chairman] presiding.

Members present: Representatives Hawkins, Martinez, Hayes, Sawyer, Payne, Poshard, Goodling, Smith, Gunderson, and Petri.

Staff present: John Jennings, counsel; Dr. June L. Harris, legislative specialist; and Jo-Marie St. Martin, education counsel.

Chairman HAWKINS. The Subcommittee on Elementary, Secondary, and Vocational Education is called to order. The hearing this morning is an oversight hearing on testing assessment evaluation to improve learning in our schools.

In order to conserve time, the Chair will not make an opening statement at this time other than to indicate that the importance of this hearing should be, viewed in terms of the importance of assessment itself. There is no way that we can achieve any of the goals in education without some form of measurement to assess where we are today or where we will be by the year 2000, or any other time.

We have invited a number of witnesses who are highly qualified in their fields. We are deeply appreciative of their participation in the hearing this morning. If other members wish to make any statement at this time, the Chair will yield to any member who desires to make a statement.

Mr. MARTINEZ. Mr. Chairman, I have a written statement that I would like submitted for the record, but I won't make statement.

Chairman HAWKINS. Without objection, the statement will be entered into the record at this point. Other statements that may be made by the members will also be included in the record.

[The prepared statement of Hon. Matthew G. Martinez follows:]

DISTRICT OFFICE
400 N. MONTEBELLO BLVD.
SUITE 100
MONTEBELLO, CA 90640
(213) 722-7731

GOVERNMENT OPERATIONS
SUBCOMMITTEE ON COMMERCE, CONSUMER AND MONETARY AFFAIRS
SUBCOMMITTEE ON HOUSING AND EMPLOYMENT
SELECT COMMITTEE ON CHILDREN, YOUTH AND FAMILIES

Congress of the United States
House of Representatives
Washington, DC 20515
MATTHEW G. MARTINEZ
30TH DISTRICT, CALIFORNIA

WASHINGTON OFFICE
U.S. HOUSE OF REPRESENTATIVES
WASHINGTON, DC 20515
(202) 225-6464

COMMITTEE ON EDUCATION AND LABOR
CHAIRMAN
SUBCOMMITTEE ON EMPLOYMENT OPPORTUNITIES
SUBCOMMITTEE ON ELEMENTARY, SECONDARY AND VOCATIONAL EDUCATION
SUBCOMMITTEE ON SELECT EDUCATION

STATEMENT FOR THE HEARING ON
STANDARDIZED TESTING IN EDUCATION

BY

HONORABLE MATTHEW G. MARTINEZ

THE SUBCOMMITTEE ON ELEMENTARY,
SECONDARY AND VOCATIONAL EDUCATION

JUNE 7, 1990

MR. CHAIRMAN, I WOULD LIKE TO SUBMIT A WRITTEN STATEMENT FOR THE RECORD.

MR. CHAIRMAN, WE ARE HERE TODAY BECAUSE SERIOUS CONCERNS HAVE BEEN RAISED ABOUT THE ADEQUACY OF TODAY'S STANDARDIZED TESTS AND ABOUT HOW THEY ARE BEING USED. // FOR EXAMPLE, IN ONE SCHOOL DISTRICT IN NEW YORK, 61% OF THE CHILDREN HOPING TO ENTER KINDERGARTEN FLUNKED A STANDARD TEST FOR READINESS./ AFTER THEY WERE ASSIGNED TO A SPECIAL TWO YEAR KINDERGARTEN, A STUDY SHOWED THAT THE TEST HAD A 50% MARGIN OF ERROR. / / THAT IS, IT WAS NO BETTER THAN FLIPPING A COIN. / IN GEORGIA THERE WERE SIMILAR RESULTS WHEN A PEN AND PAPER TEST WAS MANDATED FOR PROMOTION FROM KINDERGARTEN. / FLUNKING KINDERGARTEN IS NOT A JOKE--AND EDUCATIONAL POLICY SHOULD BE BASED ON SOMETHING MORE SOLID THAN THE FLIP OF A COIN.

THESE TESTS ARE BEING MISUSED. IT IS LIKE THE BODY-COUNTS IN THE VIET NAM WAR--WE GET HARD NUMBERS ON A WALL-CHART THAT MAKE GREAT HEADLINES BUT THEY ARE MISUSED. LIKE THE BODY COUNTS, THEY CAN TELL US WE ARE WINNING A BATTLE, WHEN WE MAY BE LOSING A WAR TO IMPROVE EDUCATION.

DEPENDING ON THE CRITIC, THESE STANDARDIZED TESTS: (A) MEASURE THE WRONG SKILLS, (B) DISTORT CLASSROOM PRACTICE, (C) FALSELY ASSURE PARENTS, OR (D) DISCIMINATE AGAINST THE UNDERPRIVILEDGED. / THE CORRECT ANSWER IS PROBABABLY "ALL OF THE ABOVE".

WE NEED TO USE TESTS WISELY TO IMPROVE EDUCATION. EVEN MORE SERIOUSLY WE NEED TO ELIMINATE THE MISUSE OF TESTS IF WE ARE NOT GOING TO SHORT-CIRCUIT EDUCATION REFORM TO PIGEONHOLD KIDS AND SEAL OFF OPPORTUNITIES FOR ALL AMERICANS.

I LOOK FORWARD TO HEARING THE TESTIMONY, AND TO FUTURE CONSIDERATION OF THIS IMPORTANT ISSUE.

THANK YOU.

ASSOCIATION FOR CALIFORNIA EDUCATION

VOLUME VIII APRIL 1990 NUMBER 4

THE COLLEGE BOARD CONTINUES EUROCENTRISM

By William Kann

The College Board, a testing service that provides the Scholastic Aptitude Tests (SATs) and achievement tests for high school students applying to college, is a private nonprofit organization. The service creates, administers, and determines the content of its tests and the subjects offered. Public and private postsecondary institutions do not. This fact tends to be lost or largely ignored in any discussion of college admissions.

Yet the fact remains that the College Board wields an enormous amount of power and is beholden to no one. This point has been brought to a head by the recent decision of the College Board to continue its policy of excluding Asian languages from its coveted ranks of achievement tests.

Postsecondary institutions such as the University of California require students to submit the SAT and achievement tests with their applications. Students may choose among a number of achievement tests, including the following language tests: Latin, German, Spanish, French, Italian, and Hebrew.

While the College Board's achievement tests create a natural demand for European languages in the high schools, Asian languages fight an uphill battle.

CATCH 22

Thus, a college-bound high school student, knowing that achievement tests in Asian languages are not offered by the College Board, will elect to take a European language in school rather than an Asian one.

College-bound students must fulfill numerous requirements to qualify for admission to a public 4-year university. There is little motivation for such students to take an Asian language and a European language. Consequently, there is little motivation for public schools to offer Asian languages if these classes will not be filled.

While the College Board's achievement tests create a natural demand for European languages in the high schools, Asian languages fight an uphill battle.

In other words, the College Board has subtly created a situation where it is not only shaping high school curriculum, it is creating a Eurocentric one.

The College Board argues that it is very expensive to develop Asian achievement tests, and even if the tests are developed, it doesn't believe that many students nationwide will take such tests. Catch 22.

ASIAN POPULATION GROWTH

The growth in the Asian population in this country, particularly in California, raises another sticky question for the College Board. By preventing bilingual Asian students from taking an Asian language achievement test, is the testing service engaging in racial bias?

The furor over the alleged Asian admissions cap at prestigious postsecondary institutions has not died down. Even UC Berkeley, a bastion of academic egalitarianism, has admitted that its admission policies unintentionally treated Asian applicants unfairly.

If a public university with all its attendant checks and balances in a city as socially progressive as Berkeley can falter, how can one expect an organization such as the College Board to be any more socially and academically responsible?

By failing to provide achievement tests in Asian languages, the College Board denies equal opportunity and equal access to Asian students who are proficient in Mandarin Chinese, Japanese, and Korean.

GLOBAL PICTURE

"Think globally, act locally." This is a saying used over and over by social activists. Yet, the phrase could easily be appropriated for this situation.

Never has the study of Asian languages been as important as it is today, with the Pacific Rim becoming an international focal point, economically, politically, and culturally.

By creating better opportunities for the study of Asian languages, the College Board and in turn our schools and universities provide a crucial first step towards a better understanding of other cultures.

Isn't it about time that Asian languages became an integral part of our educational fabric?❑

Chairman HAWKINS. The hearing will consist of a panel of experts in their particular fields. May I ask these witnesses to sit at the witness table.

Dr. Walter Haney is Senior Research Associate Director for the Study of Testing Evaluation and Educational Policy, Boston College. He's representing the National Commission on Testing and Public Policy.

Dr. Burton Faldet, President, Test Consultants, Ltd., Illinois, representing the Association of American Publishers, Inc.

Dr. Faithorn, Jr., retired business executive, volunteer teacher at the University of the District of Columbia, representing the Friends of Education, New Mexico.

Mr. Ramsay Selden, Director of the State Education Assessment Center, Council of Chief State School Officers.

Gentlemen, we will recognize you in the order in which your names have been called. May we request that your prepared statements in their entirety be entered in the record, and we hope that you will summarize or highlight your testimony so as to leave time for questioning at the end of your statements and give us an opportunity, in a very informal sense, to try to develop the subject matter which will be most productive for the committee.

We are in the process of drafting a title to an omnibus education bill, and we believe that without this title the omnibus approach to education in a more comprehensive approach would be obviously a failure if we do not, in terms of a title on assessment, develop at least the beginning of the subject.

We obviously are not going to conclude this hearing today as the only hearing on this particular subject matter. We will continue our communication with you and hope that we can call on you from time to time to help us in refining the title so that it is meaningful in terms of approaching the problem. Dr. Haney, you may proceed.

STATEMENTS OF DR. WALTER HANEY, BOSTON COLLEGE; DR. BURTON W. FALDET, TEST CONSULTANTS, LTD.; WALTER E. FAITHORN, JR., BUSINESS EXECUTIVE AND VOLUNTEER TEACHER AT THE UNIVERSITY OF THE DISTRICT OF COLUMBIA; AND RAMSAY SELDEN, STATE EDUCATION ASSESSMENT CENTER COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Dr. HANEY. Yes, sir. Thank you. My name is Dr. Walter Haney, and I'm a senior research associate at Boston College. I am here this morning representing the National Commission on Testing and Public Policy.

I'm substituting for Dr. Bernard Gifford who had hoped to be here this morning but unavoidably could not come this morning. So I wanted to first pass along Dr. Gifford's apologies for his not being here today.

What I would like to do briefly is to summarize the recent report of the National Commission on Testing and Public Policy. I have provided copies of the Executive Summary of the National Commission's report to the members of the committee. If you would desire full copies of the Commission Report, I would most certainly be

pleased to provide them. I simply could not carry copies in my luggage this morning.

Chairman HAWKINS. Doctor, we would probably need about 35 copies. Every member of the committee should be supplied with a copy.

Dr. HANEY. Thirty-five copies?

Chairman HAWKINS. Yes. If you have those available, we would appreciate it.

Dr. HANEY. I will get those sent to you as soon as I return to Boston.

The National Commission has worked for three years investigating the role of testing in the United States in both the realms of education and the realms of employment. The Commission's work was motivated by a fundamental concern that America must revamp the way it develops and utilizes human talent.

To do that in the future, as human talent is increasingly becoming the life-blood of our nation's future, we must restructure testing so that talent is promoted rather than merely screened or classified. This will require that we rethink incentives regarding educational testing and assessment.

The Commission was concerned that currently there is over-reliance on testing that is predominantly multiple choice in format and that sometimes leads to unfairness in allocation of opportunities and too often undermines vital social policies. Nevertheless, at the outset, I want to make clear that the Commission—all the members of the Commission—strongly felt that there is a vital place for testing in both our education and employment systems.

Specifically, the Commission concluded that well-designed and responsibly used assessment can be an important source of information about how our organizations and institutions are doing, what our children are learning and how well, and who among us is likely to make the most of opportunities that cannot be provided to all.

Since you have a summary of the Commission's recently released report, let me only very briefly summarize the main findings and recommendations of the Commission.

First, the Commission concluded that tests are imperfect measures with regard to both individual's learning and their employment potential.

Second, testing can result in unfairness. Some uses of testing do, in fact, result in unfairness not only for individuals but for identifiable groups of our society.

Third, in the education realm the Commission concluded that there is simply too much testing. There has been a vast increase in testing in the Nation's schools over the last 20 to 30 years, and the Commission concluded that students in our nation's schools are simply subjected to too much testing. It was estimated that students spent the equivalent of 20 million school days each year simply taking standardized tests.

If I may divert from the text findings of the Commission, let me simply illustrate some of the evidence that we accumulated to support that finding.

Chairman HAWKINS. We'll get some staff to assist you, volunteers to help out. Do you need some assistance?

Dr. HANEY. I only have a couple of charts.

Chairman HAWKINS. Okay. Well, I want to keep my staff busy. You've taken that job away.

Dr. HANEY. We had a stand for the charts, but inadvertently someone removed it just before we started. So we will substitute. But as an experienced teacher, I am quite familiar with having to improvise as I speak.

This chart simply represents the growth in state testing programs form 1950 to 1990, as summarized in an Office of Technology Assessment Report from the U.S. Congress in 1987.

It only goes to 1987. There has been an increase in state testing programs since 1987 so that now virtually every state in the Nation has a state testing program. In addition, districts have their own testing programs. Additional testing may be mandated as a result of other special programs.

Because of this repetitive testing and unclear evidence that it was providing instructional useful information, the Commission concluded that there is simply too much testing in the Nation's schools. Also, it seems clear that in addition to there being simply too great a volume of testing, that some forms of testing may in fact be undermining educational efforts in the schools. We found evidence that in many places instructional practices had been transformed simply into test preparation practices, for example.

More broadly, in the fourth finding of the Commission, the Commission concluded that testing is undermining important social policies, not just in education, but in the employment realm as well. There are several examples that the Commission cited of this general finding to illustrate this problem.

The fifth major finding of the Commission was that there's simply insufficient public accountability regarding standardized testing programs. That while tests have become instruments of public policy for maintaining accountability, there is insufficient public accountability with regard to the tests themselves.

Rarely are important tests subject to formal systematic professional scrutiny or examination in public. As a result of these general findings, the Commission concluded in its fundamental recommendation that current testing policies and practices need to be substantially restructured to help promote the development and talents of people to become constructive citizens and help institutions become more productive, accountable and just.

To help promote a vision of how this might be accomplished, the Commission made eight general recommendations which are summarized in materials I have provided so let me only briefly mention them here.

First, testing policies and practices must be reoriented to promote the development of all human talent, not just to select among people or to classify people to promote the development of all human talent.

Second, testing programs should be redirected from reliance on multiply choice tests toward alternative forms of assessment. But I wish to make clear that the Commission did not think there is any one quick fix regarding a better test or a better assessment. These sources of information about students' learning must be used flexibly and in different ways for different purposes with avoidance of

over-reliance on any one form of assessment, be it a multiple choice test or some alternative.

Third, test scores should be used only when they differentiate on the basis of characteristics relevant to opportunities being allocated. Too often a test is used simply because it is available when in fact it is not relevant to the opportunities being allocated.

Fourth, the more test scores disproportionately deny opportunities to minorities, the greater is the need to show that test measure characteristics relevant to the opportunities being allocated, because they found clear evidence that some uses of tests were in fact promoting unfairness with regard to allocation of opportunities to minorities.

The final three findings of the Commission I will summarize as follows. Test scores are imperfect measures and should not be used alone to make important decisions about individuals, groups or institutions. In the allocation of opportunities, individuals past performance and relevant experience must be considered. We can no longer tolerate bureaucratic decision-making about individuals on the basis of single test scores because of the fallibility of all test results.

Sixth, more efficient and effective assessment strategies are needed to hold institutions accountable. Right now we have considerable evidence that testing programs are providing us with misleading information about the performance of some of our vital social institutions.

Seventh, the enterprise of testing must be subjected to greater public accountability, and we must view testing for the purposes of accountability separately from testing for the purposes of promoting individual student learning.

Eighth, research and development programs must be expanded to create assessments that promote the development of the talents of all of our people. While must research and development has gone on in the past concerning testing and assessment, the Commission felt strongly that future research regarding testing and assessment needs to be motivated by the primary goal of testing and assessment to promote the development of human talent rather than simply testing and assessment to classify or measure people.

That's a summary of the Commission's report. I will be glad to answer questions and provide you with the full copies of the Commission's report as you requested. Thank you very much.

Chairman HAWKINS. Thank you, Doctor. We'll get back to you I'm sure during the questioning period.

At this point, I should like to announce that there's a vote pending in the House. Some of the members may care to go and respond to the voting or to alternate. Those who do go, I request that you return and perhaps bring another member of the subcommittee back with you.

The Chair is not desirous of going over to waste time on a useless vote such as this one.

Mr. GOODLING. I don't have an opponent this fall so I'm not worried.

[Laughter.]

Chairman HAWKINS. Well, at least we have a formal quorum and we'll continue.

Dr. Burton Faldet—I hope I'm correct in pronouncing your name—President, Test Consultants, Ltd. of Illinois representing the Association of American Publishers.

Dr. FALDET. Well Mr. Chairman, members of the committee, my name is Burt Faldet. I appreciate this opportunity to appear before you today on behalf of the Association of American Publishers.

AAP is the principal trade organization representing more than 235 member firms that publish hardcover, paperback books, professional, technical and scientific journals, computer software and classroom and education materials, including indeed tests and evaluation and scoring services.

I am President of Test Consultants, Ltd., which provides services in evaluation, design and implementation strategies to education and business. From 1965 to 1987 I was with Science Research Associates, a commercial test publisher, where I was involved in a variety of positions, management and staff, in the development, publication and use of standardized tests for schools and industry.

I've also taught some courses in measurement. I was a school psychologist, a science teacher, and director of Pupil Personnel Services.

There are several points that I would like to discuss today about the development and use of standardized tests in elementary and secondary schools from the perspective of the publisher of such tests. My statement does not address higher education, employment or military testing.

The first, and what I hope will be the most important message I'll leave with you today is that the developers and publishers of standardized tests should be seen as part of the solution for improving the quality of educational instruction, not as part of the problem.

The second message is that test diversity and competition should be encouraged to assure improved education and improved assessment instruments. Different needs for information are served by different kinds of tests. No one test can accomplish all of the diverse objectives of our educational system.

It is a serious mistake, of course, to try to make tests do what they are not designed to accomplish or to use tests as the sole means for assessment in most situations. Finally, I want to assure the committee that the test publishers working with the educational community are and will continue to expand and improve their testing products to meet continually emerging educational demands.

In the interest of time, I will leave your reading the material submitted for the record. In them, I've summarized some of our thoughts on why testing occurs from our perspective, the limits of tests, the different kinds of tests and their uses, and the role of the test developer and publisher.

Publishers are not simply printers, bookbinders and marketers. They are an integral part of the educational system, providing an essential delivery system, as well as taking the initiative for and bearing the risk of developing new and innovative materials.

What recommendations do we have for Congress? The first is that you continue to hold hearings such as this on education issues, particularly testing, as a prelude to any possible further action.

Second, Congress should continue to assure diversity of testing. No single test, no single curriculum, no single textbook can or should meet our nation's diverse educational needs.

Competition among test developers, including a vigorous private sector, should be encouraged. Publishers have a very vital role in making whatever test program may be adopted by a school work. They provide an economical and efficient delivery system for assessments of many kinds. Publishers have traditionally served as an important bridge between sound theory and sound practice.

Indeed, they have been the vehicle for getting local school acceptance of new concepts and the resulting products. They have been the primary link between those who create and those who must implement. We do not see a change in this role nor do we believe that a change is desirable. For this reason it is important to involve publishers in the early conceptualization of products resulting from sound research.

One of the crucial concerns is the proper interpretation of test results. One suggestion we would have for you might be to provide funding for targeted in-service training to teachers and administrators in interpreting test results to enable them to use tests better to improve instruction and to convey information to students, parents and the public.

State and local education agencies might be encouraged, if not required, to develop a comprehensive assessment plan which would identify instructional and accountability goals and objectives and those assessment instruments that would be used to achieve them and measure progress. The plan could include specific programs for in-service training, public information and for assuring that tests are selected, used and interpreted appropriately.

We do not believe that the Federal Government should get intimately involved in state and local testing business. Continued financial and technical support for research and development on innovative assessments, as now provided by the Department of Education and the National Science Foundation, would enable continued progress toward improving educational assessments.

Thank you for your attention. I would be pleased to respond to any questions the committee may have.

[The prepared statement of Dr. Burton W. Faldet follows:]

Association of American Publishers, Inc.

aap

1718 Connecticut Avenue, N.W. #700
Washington, D.C. 20009-1148
Telephone 202 232-3335
FAX 202 745-0694

STATEMENT BY BURTON W. FALDET
ON BEHALF OF THE ASSOCIATION OF AMERICAN PUBLISHERS
BEFORE THE SUBCOMMITTEE ON ELEMENTARY, SECONDARY,
AND VOCATIONAL EDUCATION
COMMITTEE ON EDUCATION AND LABOR
JUNE 7, 1990

Mr. Chairman and members of the Committee, by name is Burt Faldet. I appreciate this opportunity to appear before you today on behalf of the Association of American Publishers. The Association of American Publishers ("AAP") is the principal trade organization representing more than 235 member firms that publish hardcover and paperback books; professional, technical, and scientific journals; computer software; and classroom and educational materials, including tests and evaluation and scoring materials.

I am President of Test Consultants, Ltd, which provides evaluation, design, and implementation strategies to education and business. Our clients have included commercial test publishers, the American Institutes for Research, IBM, as well as individual school districts. From 1965 to 1987, I was with Science Research Associates, a commercial test publisher, where I was involved in a variety of positions in the development, publication, and use of standardized tests for schools and industry. I also have taught undergraduate courses in Measurement and Evaluation and secondary school science, and served as a School Psychologist and Director of Pupil Personnel Services.

There are several points that I would like to discuss today about the development and use of standardized tests in elementary and secondary schools, from the perspective of the publishers of such tests. My statement does not address higher education, employment, or military testing.

The first, and what I hope will be the most important message I leave with you today, is that developers and publishers of standardized tests should be seen as part of the solution for improving the quality of educational instruction, not as part of the problem.

The second message is that test diversity and competition should be encouraged to assure improved education and improved assessment instruments. Different objectives are served by different kinds of tests -- no one test can accomplish all of the diverse objectives of our diverse educational system. It is a serious mistake to try to make tests do what they are not designed to accomplish, or to use tests as the sole means for assessment in most situations.

Finally, I want to assure the Committee that test publishers -- working with the educational community -- are expanding and improving their testing products to meet continually emerging educational demands.

WHY TEST?

Measurement can be relatively exact -- but a number has no meaning until someone makes a judgment about it. That is the difference between measurement and evaluation. There are many ways to determine health; a number on a thermometer is one indicator, but it takes someone to exercise judgment as to the significance of the temperature shown, and to take the appropriate action as indicated by the reading on the thermometer. It would be imprudent, however, to rely entirely on temperature to make a diagnosis of the patient.

Why educational testing? Testing is of value to the student. It serves to provide some information that can be used by educators and parents to identify and respond to the instructional needs of individual pupils and to improve instruction of individual pupils. Testing is a means to assess progress toward specific educational objectives, as evidenced by what pupils can do in terms of skills exhibited.

Testing also serves broader, institutional goals. It assists in assessment of long-range effects of changes in the educational program, enabling comparison of (1) performance over time and to changes in the instructional program or to changes in population characteristics and (2) performance across different subject areas, such as mathematics and reading, to determine strengths and weaknesses, needs for program modification, or changes of emphasis. Testing is one means to evaluate performance for accountability purposes.

The methods of evaluating whether children are learning what is being taught have changed over the years, just as many techniques and objectives of teaching have changed. For example, standardized achievement tests and numerous other types of tests have supplemented teacher-made tests administered on a class-by-class basis.

LIMITS TO TESTING

It must be emphasized, however, that there are limits to testing. When testing is used in "high stakes" situations and results are used as a simple "pass/fail" barrier to students, or to reward or punish teachers and administrators, when the pressure becomes so intense that there is "teaching the test" rather than teaching the skills and concepts that are being evaluated, when test scores become the sole criteria for evaluating student performance or potential or the effectiveness of instruction, then testing has gotten out of hand and is being misused and abused.

Tests are a necessary but not sufficient means to assess achievement and growth in skills and abilities. What may be tested is not, and cannot be, inclusive of all of the desired outcomes of instruction.

Tests may be used as a partial basis for evaluation. Tests are concerned only with certain basic skills and abilities and are not intended to measure total achievement in any given subject or grade; they are not inclusive of all the desired outcomes of education. Standardized tests are concerned with only those areas of instruction that are amenable to objective measurement.

It should also be recognized that local performance is conditioned by many influences. The instructional effectiveness of the teaching staff is only one of these factors. Among other factors are the pupils' school and home environment, their past educational history, and the quality and adequacy of the instructional materials with which the staff has to work.

As stated in the Manual for School Administrators for one standardized test,

> At all times, the tests must be considered a means to an end and not ends in themselves. These tests have their principal value in drawing attention of the teaching staff and the pupil to those specific aspects of the pupil's development most in need of individual attention; in facilitating remedial and individualized instruction; in identifying those aspects of the whole program of instruction most in need of increased emphasis and attention; and in providing the basis for more adequate educational guidance of the individual pupil. If properly used, the results should motivate both teachers and pupils to increased, better-directed efforts in both teaching and learning.

When intelligently used in combination with other important types of information, the results obtained from these tests should prove very valuable in the appraisal of the total program of instruction. Unless they are used in conjunction with other information, however, they may do serious injustice to many teachers and to many well conceived instructional programs.

KINDS OF TESTS

Different tests have been developed to meet a variety of purposes. Some tests are subjective, both as to the matter tested and the interpretation of the results. A standardized test is an objective test that uses the same standards to measure student performance across the country; everyone takes the same test according to the same rules.

A normed-reference test (NRT) is a standardized test used to compare students' performance in terms of a carefully selected, nationally representative group, or norm, on the same test; performance is based on total test or subtest scores. (In contrast, for some tests, such as the SATs and ACT, the norm is based on the others taking the test, rather than to a standardized national norm.)

A criterion-referenced test (CRT) differs from a normed-reference test primarily in how test scores are interpreted and used. A criterion-referenced test is used to evaluate and report performance in terms of specific instructional objectives or skills, stated in measurable terms.

These labels are not mutually exclusive. Many criterion-referenced tests are normed, and many norm-referenced tests may be subject to criterion-referenced, content-based interpretations.

Teacher-made tests generally are intended to provide information about individual student's performance on specific, classroom-oriented, curricula or specific needs for information about students. These tests are frequently supplemented by textbook tests, which are developed by textbook publishers and may appear in textbooks or be provided to teachers as supplementary instructional materials. Both of these tests are associated frequently with grades on report cards and help measure a student's progress in class, as well as facilitate individualized instruction.

Tests can also be in a variety of formats. Multiple-choice tests offer the advantages of objectivity and uniformity of

scoring, ease of administration and scoring, and low cost. There are disadvantages to such tests, particularly if they are utilized as the exclusive method of assessment. "Performance-based tests," "authentic assessments," or "alternative assessments" generally are open-ended tests that are not multiple-choice. They include essays, writing samples and portfolios of work, practicums, or oral or visual demonstrations. They generally are more expensive, labor-intensive, and require more training and preparation to administer and evaluate -- factors which also can make them affirmative educational tools. The same concerns for validity and reliability, standardization if used for comparisons, and abuse if used in high-stakes situations that are raised with multiple-choice tests are applicable to performance tests.

Performance testing and standardized testing are not mutually exclusive. It is important to point out that for several years writing and listening assessments -- performance tests -- have been offered by test developers as part of their standardized test batteries. Publishers are now offering portfolio tests to supplement their current test batteries.

What are the particular advantages of a norm-referenced, standardized test? It ensures reliability and validity in data collection, analysis, and interpretation. It enables evaluation of student achievement in various grades and subjects for the purpose of aggregating and reporting achievement gains in terms of a common reporting scale (e.g., normal curve equivalent or grade equivalent), with nationally representative norms. It provides an objective, rather than a subjective, assessment.

Norm-referenced, standardized tests also enable identification of problems in specific skill or subject area deficiencies for teacher attention and remediation. This may be particularly important in the early grades.

Norm-referenced, standardized tests use the same or parallel test items for all students, which makes scores for all students comparable; use of one level per grade facilitates criterion-referenced interpretation of results for classes, buildings, and systems. Individual scores can be related to comparable national norms. One skill can be compared to another on a pupil, class, building, or system basis.

A classroom may have such a wide range of skills that no simple test can be equally suited to the entire range of achievement; NRTs for different levels of achievement can be administered so that each pupil takes the level that corresponds most closely to the individual instructional objectives and levels of skill development.

ROLE OF THE TEST DEVELOPER AND PUBLISHER

Test developers adhere to strict standards, as developed by the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education in the Code of Fair Testing Practices in Education, a copy of which is submitted for inclusion in the record. Demonstration of reliability and validity also must be provided to test users, showing that the test meets its intended purpose and its appropriateness for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested. Several books give in-depth, candid reviews of available tests, include the Mental Measurement Yearbook, published by the Buros Institute of Mental Measurements, while guides and evaluations are published by the ERIC Clearinghouse on Tests, Measurement, and Evaluation and by other organizations.

Standardized tests generally are professionally developed tests distributed by commercial test publishers; development may be by the publisher, educators or other non-profit organizations (under royalty or other forms of compensation), or by governmental entities alone or in cooperation with publishers (such as under the National Science Foundation's "Publisher Initiative").

The role of the commercial test publisher in test development is very extensive. Based on information from a variety of sources, including the educational community, the test publisher determines if there is a need for a test and whether it will be financially viable. If the answers are in the affirmative, a decision is made as to the type of test to be developed, i.e., a norm-referenced or criterion-referenced test, or a combination of the two. In addition to the type of test, the format also must be determined. Publishers also respond to test requirements of state and local education agencies.

Extensive research is required for "building" a new test or revising an existing test. Test items are written by educators and professional test item writers. They are selected after extensive research on educational objectives; curriculum; goals, objectives, and standards; textbook and instructional material content; and what is to be measured and how. Error-free items must be developed that will withstand the scrutiny of hundreds of thousands of teachers and students over a long-period of time. Vocabulary and readability levels must be appropriate for the students to be tested. Items must also be free from ethnic, gender, or cultural bias.

At least one tryout to obtain data for standard item analysis and summary test statistics is needed. This data is

used to select items with desirable characteristics. Typically, an experimental edition will contain at least twice the number of items required for the final test, to enable the publisher to reject undesirable items and still retain a sufficient number of items for a final test of suitable length. Items for a norm-referenced test will be rejected if too many examinees select the answer. In a criterion-referenced test students are classified in terms of mastery/non-mastery, so items will be selected that will have a large number of correctly-selected answers.

Experimental test items are reviewed by educators and curriculum specialists and are then field tested with large numbers of student to check their responses. The comments of the reviewers and the data generated by the field test are used to select the items for the final edition of the test.

The final, or standardized, version of the test is administered to carefully selected groups of students whose characteristics are similar to those of students throughout the nation. The information obtained is then aggregated into norms so that individuals tested in the future may be compared to the original national sample. This is the process of standardization, and the normative information obtained from the process is crucial to educators, parents, and students. Without it, there would be no way of knowing how a single score on a specific test compared to the scores of other students in the nation.

Publishers develop guidance materials to assure that the final test is administered in accordance with the standardization, and to provide instruction on how the test is to be interpreted. Information is also developed and provided on the technical characteristics of the test to support its reliability and validity.

Scores can be reported and evaluated in a multitude of ways, for different uses. Rather than trying to describe scoring and interpretation in my testimony, I am submitting for the record an excerpt from Understanding Achievement Tests: A Guide for School Administrators, published by the ERIC Clearinghouse on Tests, Measurement, and Evaluation, on "What Types of Test Scores Are There."

Much controversy has been generated recently over norm-referenced testing. To address these concerns, I am attaching to this statement several articles from commercial test publishers that were included in the Summer 1988 Educational Measurement: Issues and Practice that provide an extensive review of these issues.

WHAT SKILLS ARE TESTED?

Higher order skills, not just basic skills, can be measured, even in a multiple-choice format, in a standardized test (remembering that it was only a very few years ago that publishers had to respond to demands for assessment instruments for the "back to basics" movement). We recognize that there are more direct ways of measuring higher order skills.

As previously stated, the multiple-choice format used in assessment instruments has some attractive features. It is an efficient and effective way of measuring many educational objectives. While we recognize that it has limitations as well, it is important to recognize that most measures, including criterion-referenced and performance tests, are samples of behavior from which inferences can be drawn. For example, a multiple-choice mathematics test, which includes five exercises in addition of two-digit numbers with carrying, is a sample of all the possible two-digit numbers that we want a student to be able to add. For efficiency, we chose five exercises, and based on the student's performance on those, we infer what the student could do if presented with many more. Similarly, we may present a situation with several complex problem-solving exercises in a multiple-choice format. Based on performance, we can make some inferences about the student's performance in some of the higher order skills in the mathematics area.

Similarly, we can infer some important aspects of performance in writing from items commonly presented in multiple-choice language arts tests.

Neither the problem-solving nor language arts tests are substitutes for direct observation of student performance over time and in different situations in solving problems and in producing written material.

Reiterating a constant theme of this statement, that tests need not be mutually exclusive, I again want to point out that publishers of standardized tests currently also offer tests of listening skills and writing in addition to multiple choice tests, as well as portfolio tests.

Whether multiple-choice or performance tests, the keywords for the future, as they are today, are validity and reliability. Publishers cannot and should not market a test unless it has been demonstrated to be valid and reliable. This requires time and money, extensive research and development, testing and reworking to assure that the test works.

RECOMMENDATIONS AND SUGGESTIONS FOR FEDERAL POLICY

On behalf of the publishers of standardized tests, I welcome this opportunity to meet with the Committee and discuss standardized tests and our role in the educational process. As I said at the beginning of my statement, publishers want to be part of the solution, not part of the problem. Publishers are not simply printers, bookbinders, and marketers. They are an integral part of the educational system, providing an essential delivery system as well as taking the initiative for and bearing the risk of developing new and innovative materials. Just as Congress would not think of addressing the future of the automobile without consulting with automobile manufacturers, publishers should continue to be consulted and included in your continued deliberations over the quality of education.

What recommendations do we have for Congress? The first is that you continue to hold hearings such as this on education issues, particularly testing, as a prelude to any possible future action.

Second, Congress should continue to assure diversity of testing. No single test, no single curriculum, no single textbook, can or should meet our nation's diverse educational needs. Competition among test developers, including a vigorous private sector, should be encouraged.

Publishers have a role in making whatever testing program that may be adopted by a school work. They provide an economical and efficient delivery system for assessments. Publishers have traditionally served as an important bridge between sound theory and sound practice. Indeed, they have been the vehicle for getting local school acceptance of new concepts and the resulting products, and for enhancing and modifying those products as needed. They have been the primary link between those who create and those who must implement. We do not see a change in this role, nor do we believe that a change is desirable. For this reason, it is important to involve the publishers early in the conceptualization of products resulting from sound research.

One of the crucial concerns is the proper interpretation of test results. One suggestion might be to provide funding for targeted, in-service training to teachers and administrators in interpreting test results to enable them to use tests better to improve instruction, and to convey information to students, parents, and the public.

State and local education education agencies might be required to develop a comprehensive assessment plan, which would

identify instructional and accountability goals and objectives and the assessment instruments that would be used to achieve them and measure progress. The plan could include specific programs for in-service training, public information, and for assuring that tests are selected, used, and interpreted appropriately.

We do not believe that the federal government should get into the state and local testing business. Continued financial and technical support for research and development on innovative assessments, as provided by the Department of Education and the National Science Foundation, would enable continued progress toward improving educational assessments.

I would be remiss if I did not point out that while publishers are trying to respond to the need to develop challenging and innovative tests (parallel efforts are being undertaken by publishers of textbooks and other instructional materials), federal tax policy is frustrating its achievement.

The Department of the Treasury is insisting that publishers of tests and instructional materials capitalize research and development and other pre-publication costs, a position that falls with special weight on preparation of new tests and instructional materials, with their high development costs, high risks, and long lead times. This approach is shortsighted as a matter of educational policy because it discourages the development of the innovative quality tests and textbooks our schools need. It is also discriminatory and unjustified tax policy because it requires capitalization of product development and research costs that, for any other industry, could be deducted in the year incurred. We have requested the tax-writing committees (and the Administration) to provide appropriate relief, but the outcome remains very uncertain. This Committee's assistance in assuring that tax policy does not frustrate education policy would be most welcome.

Thank you for your attention. I would be pleased to respond to any questions the Committee may have.

1779h

EDUCATIONAL MEASUREMENT
ISSUES AND PRACTICE
Volume 7, Number 2
Summer 1988

Riverside Comments on the Friends for Education Report

Edward C. Drahozal
Riverside Publishing Company
and
David A. Frisbie
The University of Iowa

The authors, both affiliated with Riverside Publishing Company, discuss the factors they think explain the Lake Wobegon phenomenon and call for more appropriate use of normative comparisons and more complete reporting of test results.

The final report of Friends For Education (FFE), "Nationally Normed Elementary Achievement Testing in America's Public Schools," has appeared in several preliminary forms, and portions of its contents have been discussed in several national newspaper reports. The following comments are based solely on the content of the final report, and, though we take exception to some of the data analysis methods employed, some of the interpretations offered, and some of the conclusions reached, these remarks focus primarily on questions raised by the report. These are issues that complicate the use of aggregated test scores—with regard to both their formation and their interpretation—to synthesize state versus national comparisons in the face of different test batteries, different standardization samples and methods, different years of norming, and different score scale units for reporting.

The study of state and district performance reported by FFE appears to have been conducted as carefully as possible under the circumstances. The issues that the report raises—accuracy and comparability of norms, currency of norms, selectivity of pupils tested and reported in reports to the public, and temptation to teach specific content when educators are under accountability pressure—are not new. What comes as a genuine, unexpected, disappointing shock is the apparent universal appeal of the

Edward C. Drahozal is Senior Program Administrator at The Riverside Publishing Company, 8420 Bryn Mawr Avenue, Chicago, IL 60631. He specializes in practical aspects of interpretation, use, and reporting of elementary and secondary tests, primarily standardized tests.

David A. Frisbie is Associate Professor and Assistant Director of the Iowa Basic Skills Testing Program, University of Iowa, 316 Lindquist Center, Iowa City, IA 52242. He specializes in achievement testing.

simplistic objective of being above the national average and the extent to which schools are successful in managing somehow to appear above the national average when faced with pressures and even ultimatums from politicians, press, the courts, and even watchdog groups.

Despite the shortcomings that can be cited regarding the nature of the data that FFE analyzed and reported, there is ample evidence to warrant close examination of the group's fundamental question: Why are so many pupils or schools or states appearing to perform above the national average? The question seems as simple and straightforward as the one posed several years ago: Why are test scores declining from year to year? We believe that the question raised by FFE rivals the score-decline question in significance and, as was true of the score-decline inquiry, this search for resolution is likely to yield multiple, concomitant explanations. There is no single best answer. A closer examination of the issues by FFE, the publishers, and the state and district test coordinators might enhance our ability to use test data to further our primary goal—to improve the quality of instruction provided in our schools. With this purpose in mind, the remainder of this paper is devoted to identifying what we believe are the most crucial issues and to presenting a scheme that we would use to compare the performance of state or district groups with national pupil or school norm groups.

Some Major Issues

Accuracy of Norms

National norms for standardized achievement tests are based on a sample of pupils and schools (attendance centers) obtained through a complex, multistage sampling scheme. Each publisher strives to ensure that the national population of pupils and schools is properly represented in its norms sample. For example, in the standardization of the Iowa Tests of Basic Skills (ITBS) in 1984–85, districts were chosen on the basis of geographic region, enrollment size, and socioeconomic characteristics of the communities served. The standardization is a joint responsibility of the authors, publishers, and school personnel. Rigid conditions for district participation included the provision for sampling attendance centers of the district by the publisher rather than by the school administration. An adequate sampling plan is necessary but not sufficient to guarantee adequate norms. Only if the plan is realized, only if the sample obtained reflects the sample desired, will the norms represent national pupil or school achievement accurately.

To the extent that any publisher's norms misrepresent the national distribution of pupil and school achievement, comparisons with either of these norm groups will distort the estimated achievement level of the group in question. An underrepresentation of high-achieving schools or high-achieving pupils will cause the national norms to be "softer" than they ought to be. That is, an average-achieving pupil will appear to be above average when referenced to a group whose average is below their theoretical or "true" average.

The sampling plan, nature of the obtained sample, and weighting schemes used in the standardization of each achievement battery in question should be examined to determine the representativeness of the published norms. This should be done separately for pupil and school norms.

Recency of Norms

It is a well-documented fact that achievement in grades 3–8 has been rising steadily since the late 1970s. Though the year-to-year differences might be regarded as minor (.3 of a grade-equivalent month, on the average), the cumulative effect over 10 years is significant (approximately 3 months, on the average). Obviously, those who compare the 1987 performance of their pupils with that of other pupils who were tested in 1978 (national standardization) will be using "softer" norms and will have more pupils appearing to be above the national average than really are.

We have published information on changes in student performance for the past 30 years. Data for 1955 to 1984 are summarized on pages 148–153 of the new ITBS *Manual for School Administrators* (Hieronymus & Hoover, 1986). Differences in performance vary by test, grade, and score level. The 1977–85 composite score differences are eight to nine percentile ranks (PRs) at the median in *most* grades, but differences in language exceed 10 PRs in several grades at several score levels.

In periods of fluctuating achievement levels, the recency of the norms is a critical issue. When achievement levels are relatively stable over time, as they have tended to be at the grade K–2 levels, "old" norms do not interfere with score interpretations, assuming that we have curriculum stability as well.

Nature of Tested Population

If we have good reason to believe that pupils in a given state should have scores, on the average, below the national average, we must be certain to define the population for which we expect the prediction to hold. There are several related issues regarding this point with respect to the FFE data. If State X reports a mean normal curve equivalent (NCE) for 45,000 fourth graders, we should ask these questions: How many fourth graders were tested but not included in the computation of the mean, and what is the nature of the scores of those who were excluded from reporting? How many fourth graders are there in State X who were not tested and, consequently, who were not included in the reported scores? And what are the achievement levels like for these students who were not tested?

Based on the Department of Education's Center for Education Statistics fall 1985 enrollments projected to 1986, the percentage of students for whom scores are reported in the FFE report varies from a low of about 85% to more than 95% of total grade enrollments for *most* states for which full-grade testing was reportedly done. (For one state with public school enrollments of about 48,000 students per grade, averages and PRs are reported for approximately 37,500 students, which is about 80% of the total enrollment.) The discrepancy between the reported state scores and the expectations in the FFE report may be in part due to such differences between tested and total enrolled populations of students and

specifically to the nature of the portion of the student population not tested.

Adequacy of Expectations

Educators have developed some expectations about how pupils and groups of pupils might perform on achievement tests based on their study of the relationship of school achievement to other social, political, and economic variables. This is why we use such variables as enrollment size and socioeconomic status for stratified sampling in standardizations. FFE has used some of these relationships in attempting to develop expectations for state-level and school-district-level performance. Per-capita income, graduation rate, and college entrance score averages are among the "standard barometers of excellence" employed by FFE. Though we do not deny the value of these indicators as part of the prediction equation, we realize that it is not possible to predict achievement in this way with high accuracy. For example, the achievement test performance of Iowa pupils is among the very highest in the nation, yet these facts about educational conditions in Iowa seem inconsistent with that high level: Iowa ranks 27th among states in per-pupil expenditure, 39th in average teacher salary, and 44th in spending increase from FY86 to FY87.

In view of the less-than-perfect relationships between achievement and these other variables, the precision of whatever expectations about achievement we may formulate should be tempered. That is, what we are able to say with reasonable assurance about how many pupils or schools should score above a specific point (the mean in the national norms distribution) is not very useful. Consequently, we might instead settle for statements like these for State X: "About 40% of the fifth graders tested should score between the 25th and the 75th percentiles on national pupil norms," or "About 49–55% of the third graders tested should score above the national pupil median (50th percentile)." Of course, the ability to make such statements depends on a far greater understanding of the statistical relationship between those variables than most states probably have been able to determine.

Teaching the Test

Pupils and their teachers who participate in the standardization of an achievement battery have not had an opportunity to see or study the specific test questions used. Thus, having no practice on the specific test questions is one of the stringent preconditions of the standardization process. Subsequently, when these norms are used to interpret the scores of pupils who have been drilled with the exact test questions, the result is an overrepresentation of the amount of knowledge and skill possessed by such pupils. Likewise, when the scope of the curriculum is narrowed to encompass primarily the objectives measured by the exact test questions, the relative standings of the pupils who experienced the restrictive program of study will be overestimated.

No publisher condones this use of tests, and few teachers probably follow such abominable practices. Those who do are nearly always motivated by . . nificant negative consequences associated with scores that might turn out to be below expectation (not always synonymous with national average). Unfortunately, for some educators, job retention and salary increases are tied directly to the test scores of their pupils. The authors of the ITBS have always decried the use of achievement scores for such purposes and instead have campaigned for the use of these scores to improve instruction directly.

If certain tests are to be used strictly for accountability purposes, their security must be ensured so that the scores that result will be valid for that purpose. The dollars required to assure states and districts that the test forms they will use are secure would be far greater than the value of the information derived from using the secure forms. Those dollars would likely have greater and more visible impacts on learning if devoted to direct instruction instead.

Score Analysis and Interpretation

With which norm group, pupils or schools (attendance centers), should averages from State X be compared to interpret the scores of pupils from that state? With which norm group, pupils or schools, should averages from District A be compared? There are only two choices, pupils and schools, because no publisher provides norms for school districts or for states. This is a fundamental issue currently facing the Council of Chief State School Officers as they contemplate options for providing for state-by-state achievement comparisons in the future. The choice to be made is not a matter of personal preference but a matter of the logical correspondence between the units to be compared. That is, averages of school buildings should not be referenced to a distribution of individual pupil scores, district averages should not be referenced to the distributions of either school building averages or pupil scores, and state averages should not be referenced to any of these three distributions. In view of the differences between these separate distributions, it is most logical to reference a score or average score to its own kind. When the most logical referencing is not possible, appropriate caution should be exercised.

The national pupil norm group includes pupils whose scores on a test are as high as perfect (PR = 99) to those whose scores are as low as zero or chance average (PR = 1). No school (building or attendance center) is likely to have an average score that is perfect or zero. In fact, on the ITBS and any other test with recent school norms it is reasonable to expect that no school will have a raw or scale score average higher than PR 88 or lower than PR 12 compared to the pupil distribution. Because many school districts are single-grade-within-single-building entities, the distribution of school district averages probably would encompass the same range as the distribution of school building averages. The school district distribution, however, is likely to be markedly more leptokurtic and less variable than the school average distribution. In terms of the pupil distribution, the distribution of district averages might range, *effectively*, between PR 75 and PR 25. Finally, *most* of the state averages on a test for a given grade might well have actual bounds that corre-

spond to PR 60 and PR 40 on the pupil distribution.

Because norms for district averages or for state averages are not available, districts and states often use the pupil and school norms that do exist. When a district average is referred to the pupil norms, it should be thought of as the score obtained by the average pupil in the district. We might find, for example, that the average pupil in District A scored higher than 63% of pupils nationally. Using the same rationale and the estimate given above, the average pupil in most states is not likely to exceed PR 60 or fall below PR 40. The value of such information is highly questionable.

A matter related to this general issue of analysis concerns the methods of computational precision used to aggregate and convert scores. As an example of the problem, a grade 4 school average GE composite score of 42.0 (obtained in the fall) on the ITBS has a PR of 46, and a score of 43.0 has a PR of 53. By interpolation and rounding, an average GE of 42.5 corresponds to a PR of 49.5 or 50. If GEs are rounded before converting to PRs, a 42.5 could be treated as a PR of 46 or 53, depending on the rounding convention adopted. Of course, this illustration underplays the magnitude of the distortion that could result with distributions of either school district or state averages.

Publisher-User Responsibilities

Though it is in the best interest of both publishers and test users to have tests and scores used properly, neither can ensure that the other will do its part willingly and unselfishly. Publishers must be counted on to standardize and analyze results in professionally acceptable manners. They must guard against potential misuse by informing educators of the intended uses of the tests they publish and warn against the possible misuses that might be anticipated. Publishers must do their utmost to provide test materials only to those who are at least minimally qualified to handle the tests and scores in a professional way. State directors, superintendents, teachers, school boards, and the public, generally, do not have the resources to monitor the effectiveness of publishers in attending to these obligations.

Publishers, on the other hand, cannot monitor the use of their instruments effectively to curtail misapplication, misuse, or misinterpretation. Often after the fact, a publisher can recognize inappropriate use—whether intentional or unintentional—and attempt to persuade the user to modify a proposal or report. Some school districts perform extensive audits to ensure that all students who were to be tested in each attendance center were actually tested. Some districts also audit results and retest suspect groups. But for the most part, publishers are not aware of and have no control over school districts' test administration conditions, the students included in summary data reported to the public, or methods used to synthesize data to make test results more palatable for less sophisticated consumers.

Most test authors and publishers go well out of their way to comply with the standards for educational and psychological tests adopted by the profession. Test score users—teachers, administrators, legislators, and other public groups—tend to know far less than they should about the nature of tests or the principles with which test makers intend for scores to be used. We should not denounce a test because a state committee uses the wrong norms or incorrect statistical analysis procedures in reporting. Likewise, we should not blame users for results based on shoddy standardization procedures or on inadequate or deceptive descriptions of such procedures.

Finally, publishers are obligated to clients to maintain the confidentiality of test data. It has been and should continue to be each client's decision to release test data and to determine the nature of any data to be released. Reporters, citizens, citizens' groups, and others who wish to obtain test data should respect this publisher-client relationship and seek release from the school district or state, depending on their level of interest and the dictates of state law.

TABLE 1

Percentages of State X Pupils Performing Within Selected National Pupil Percentile Intervals

National percentile rank	National percentage	Grade K	1	2	3	4	5	6	7	8	Average K-8
90-99	10	20	21	24	20	21	20	20	21	19	20 7
75-89	15	22	24	23	24	23	25	24	24	23	23 5
50-74	25	26	29	25	27	28	28	29	28	28	27 4
25-49	25	19	18	18	17	16	17	17	17	19	17 6
10-24	15	8	6	7	7	8	6	7	6	7	6 9
1-9	10	5	2	3	4	3	3	3	4	4	3 4
Percentage above national median		68	74	72	71	72	73	73	73	70	72
Percentage below national median		32	26	28	29	28	27	27	27	30	28

A Sample Reporting Method

We recommend an approach like the one described below for states that wish to describe the achievement levels of their pupils in relation to pupils in a nationally representative norm group. Exactly the same procedures could be used with school building (attendance center) data. Table 1 shows national PR ranges in the first column and the corresponding percentages in the second column. The body of the table shows, separately for each grade, the percentage of pupils in State X that obtained national PRs in each range. The last column shows the row averages of the percentage values. (Note that these are percentages and not percentile ranks and, consequently, it is acceptable

to average them.) The bottom two rows indicate, again by grade, the percentage of pupils above and below the national median. A histogram with one distribution superimposed on the other or a simple bar graph would provide a helpful visual display of the same information.

The main advantage of this method of reporting compared with reporting simply the percentage scoring above the national median is obvious. Between-grade differences and similarities can be examined, but most important, discrepancies from the national distribution can be accounted for in each of several segments of the distribution. If all we know is that 72% are above the national median, we do not know if the "extra" 22% are mostly located very near the median, mostly spread through the upper half, or mostly concentrated in the tail. Also, we do not know if the extra 22% are shifted from the lower tail, from throughout the lower half, or from just below the median.

Many districts use a reporting procedure similar to the reporting scheme described above. We recommend that such tabular data be supplemented with at least the following sorts of information: testing date, test form and level(s) used, type and date of the norm used, and percentage of eligible students tested.

Riverside Publishing Company and its representatives do not believe that the average pupil in every state has scores above the national median on the ITBS. We are confident in our standardization procedures and have subjected those procedures to public scrutiny in detail in the *Manual for School Administrators*. We have updated our norms at least every 7 years and, when achievement showed a pattern of increase in the early 1980s, new norms were obtained even though new test forms were not introduced. We are making plans to provide annual national norms updates for next year. Our manuals caution users about appropriate use of norm groups for varying purposes. Our hope is that the issues raised above will cause FFE and state and district test coordinators to reassess their analysis and reporting procedures to ensure that conclusions reached are based on a valid foundation rather than data of questionable origin and manipulation.

References

Hieronymus, A. N., & Hoover, H. D. (1986). *Manual for school administrators: Iowa Tests of Basic Skills*. Chicago: Riverside.

A Response to John J. Cannell

Joanne M. Lenke and John M. Keene
The Psychological Corporation

Two representatives of The Psychological Corporation present their reactions to the Cannell report and call for better explanations for the public of the meaning and limits of norm-referenced scores

In recent years, public attention has focused on standardized achievement test results. These results, which are intended to describe the performance of individuals in relation to one another, are now often used to describe the performance of groups of students. In a report entitled "Nationally Normed Elementary Achievement Testing in America's Public Schools: How All 50 States Are Above the National Average," John Jacob Cannell attempts to cast doubt on the validity of the information being reported to describe the achievement of students as aggregated at the state and/or district level. The report states, "These tests allow all the states to claim to be above the national average! The tests . . . allow 90% of the school districts in the United States to claim to be above average. More than 70% of the students tested nationwide are told they are performing above the national average."

In response to Cannell, it is fair to say that many states and school districts report above-average performance in reading, mathematics, and/or language in the elementary grades. We do not believe that this is an attempt to misrepresent students' achievement in the nation's schools. Let us examine three very

Joanne M. Lenke is Vice President, Measurement, at The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204-2498. She specializes in test development and norming, scaling, and equating tests.

John M. Keene is Director, Admissions Credentialing, Customized Measurement, and Research, at The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204-2498. He specializes in educational measurement.

important issues related to the interpretation of this information: (a) group performance relative to a national norm, (b) local performance relative to national performance, and (c) the stability of achievement test norms over time.

Interpreting Group Performance Relative to a National Norm

When a test is standardized, or normed, the test is typically administered to hundreds of thousands of students nationwide. This norming sample is drawn to reflect specified demographic characteristics of children attending school in the United States. Such demographic characteristics include socioeconomic status, ethnicity, region of the country, and size of school district. Percentile ranks are then derived from frequency distributions of individual students' scores at each grade. Norms provide a mechanism for describing a student's performance relative to that of other students in the same grade from across the country at a particular point in time.

The use of these norms to describe group performance must be interpreted carefully. For example, if a state's average score in reading is at the 54th percentile, the proper interpretation of this score is that the average, or typical, *student* in the state performed better than 54% of the norming sample. It is not appropriate to conclude that all students in the state are above average in reading, that the state as a whole is above average in reading relative to other states, or that the state as a whole is above average in reading relative to the national norm.

The approach used by some states and school districts in the reporting of group performance is to report the percentages of students scoring, say, "at or above the 50th percentile," or "in the average and above-average range." Although this method of reporting is appropriate because it maintains the relationship between individual performance and the national norms, the reported percentages should be accompanied by corresponding percentages for the national norming sample. Although it is obviously the case that 50% of the national sample of students scored at or above the national median at the time the test was standardized, it may not be the case that 50% of the national sample scored at or above the national mean raw score or national mean scaled score. If the reporting metric is something other than the percentage of students scoring at or above the national median, the appropriate national comparison should be provided so that proper inferences about the data can be made.

Interpreting Local Performance Relative to National Performance

It is unlikely that the demographic characteristics of the students in *any* state or school district mirror those of the nation as a whole; it is equally unlikely that the curriculum of any state or local district is as diverse as that of the nation as a whole. Furthermore, it is not necessarily the case that the guidelines set forth by the test publisher with regard to the testing of handicapped or limited-English-proficient students in a norming program are the same as those used in actual practice. If there were a state or district whose demographic characteristics matched those of the nation, whose curriculum was as diverse as that of the nation as a whole, and whose administration guidelines and procedures were consistent with those used by the publisher for the norming sample, one would expect the average student in the group to score at about the 50th percentile. To the extent that differences exist, we must remind ourselves that when local group summary scores are interpreted in reference to a national norm, the interpretation has to be placed in the proper context, simply that of the group's average student relative to the national norm. Because it is unlikely that the students tested in any given state or district are typical of the nation in all respects, it would be unreasonable to expect *any group* to be *at* the national average.

Test purchasers, districts as well as state agencies, often select tests through a process that examines the match between the test content and the local curriculum. In many cases, the selected test is the one that best reflects the local curriculum. Test users selecting tests on this basis may have an advantage over the norm group because the test is likely to be more valid for assessing performance in the local curriculum than it is for assessing the performance of a national sample of students being exposed to different curriculums, presumably having somewhat different emphases.

The Stability of Achievement Test Norms Over Time

Cannell's report suggests that the use of "old" norms is partially responsible for high achievement test scores. Presently, test publishers produce new editions of their tests on a 7-to-9-year cycle, and current norms are provided with each new edition. Because test adoption cycles do not necessarily coincide with test revision cycles, it is conceivable that the norms for a newly adopted test may be 2 or more years old. Therefore, it is critically important that empirical norming dates accompany the reporting of achievement test results.

It is very encouraging to note that today's students are performing better than their counterparts did in the late 1970s and early 1980s. Evidence of this improvement in performance can be found not only from research that test publishers have conducted in equating newly published tests to previous editions, but also from a recent research study conducted by The Psychological Corporation with the current edition of the *Stanford Achievement Test* Series. First standardized in the 1981–82 school year, the Stanford Series was administered to a nationally representative sample of 350,000 students in spring and fall 1986. The sample was further stratified according to "user" and "nonuser" groups, where "users" were defined as school districts that had been using the Stanford in one or more grades for at least one year in their districtwide or statewide assessments. The results of this study revealed that "users" outperformed "nonusers," and, more importantly, that "nonusers" performed better than the original norming sample in mathematics, reading, and the language arts in the elementary grades. Two important generalizations can be made from this research. First, test scores do tend to increase when the same test series is used year after year. However, this should not necessarily be attributed to "teach-

ing to the test"; rather, the test results provide a needed focus on areas in need of improvement. Second, educational achievement did improve from 1982 to 1986 in some subject areas in the elementary school grades. Therefore, more current norms for the Stanford Series have been developed and are available to users of the battery.

During this time of educational improvement, it is important not to lose sight of the fact that use of the same norms over a period of years enables the test user to demonstrate improvement relative to a constant reference group. Even if it were economically feasible for test publishers to produce representative national norms more often, frequently updated norms represent a "moving target," where educational gains (or losses) would be masked by the relative nature of the information. The level of achievement of students in the United States has increased in recent years, and educators must have the opportunity to demonstrate these gains in order to ensure the necessary support of the local community in improving the quality of education. The education of young people must continue to improve, and norm-referenced achievement tests are useful tools in this endeavor.

Conclusion

Because the public is expecting norm-referenced scores to represent standards of performance, we, as educators, must assist the public in becoming better informed about the interpretation of test results. National normative data provide extremely important information for making sound educational decisions. The degree to which these decisions are defensible depends on a clear understanding of the strengths and limitations of the data.

The Time-Bound Nature of Norms: Understandings and Misunderstandings

Paul L. Williams
CTB/McGraw-Hill

Presenting a view from CTB/McGraw-Hill, the author discusses the time-bound nature of test norms and argues that the phenomenon of 'most elementary students' scoring above averages from previous years' norms is a result of generally increasing levels of achievement.

Recent interest in the topic of the time-bound nature of normed scores has resulted, in part, from allegations made in a report issued by the Friends for Education. The key element of the argument put forth in the Friends for Education report is that too many students appear to exceed the national average. Data have been presented in the report which are said to show that more states and school districts are scoring above average than one might initially expect.

It is an interesting phenomenon that it is through the vehicle of the Friends for Education report that the time-bound nature of norms has received some measure of public attention. The fact that norms have always been referenced to the year of test standardization is something that has been so universally known and understood by testing professionals that it has not had a large measure of attention focused on it. Perhaps that will prove to be an important singular contribution of this issue of *Educational Measurement: Issues and Practice*.

The Cyclical Nature of Test and Norms Development

The evolution of norm-referenced tests (NRTs) as valuable assessment instruments has been characterized by the expansion of the purposes for testing. In the earlier versions of NRTs (in the mid-1960s to the mid-1970s), the primary purpose was to provide accurate normative scores so that group and individual comparisons could be made to a national profile of achievement. Using this information, school adminis-

Paul L. Williams is Director of Research and Measurement at CTB/McGraw-Hill, 2500 Garden Road, Monterey, CA 93940. He specializes in educational testing and measurement.

trators could evaluate programmatic and individual strengths and weaknesses so that appropriate instructional intervention and resource allocation could be applied. Additionally, using multiple-year testing, longitudinal trends in achievement could be monitored.

An expansion of these purposes took place with the publication of the *California Achievement Test* (CAT), *Forms C and D* (CTB/McGraw-Hill, 1977). This test battery, for the first time, allowed scores for instructional objectives to be reported from an NRT for individual examinees. Although earlier NRT test versions did allow test administrators to use item analyses for minimal diagnostic purposes, CAT C and D provided specific instructional objective scores for the purpose of more individualized instructional planning.

The schedule for the publication of norm-referenced tests has followed a basic, industrywide cycle of between 5 and 8 years for the same test series. In the instance where a test company has more than one NRT series, such as CAT and the *Comprehensive Test of Basic Skills* (CTBS), publication is staggered so that one test of the series is published about every 3 or 4 years.

This cycle has been dictated by several factors. The first factor has been the speed with which curricular changes take place in the nation's schools. NRTs are designed to reflect the predominant achievement outcomes and curricular trends in the nation's schools. When a new form of an NRT is developed, content considerations are of paramount importance. Although curricular trends have a major impact on the content of NRTs, these trends do not change so fast in the schools that more frequent revisions of a test series would be justified based solely on them.

At the time an NRT is revised, the collection of data for the generation of new national norms takes place. Using a national probability sample, data are collected for several hundred carefully selected school districts and hundreds of thousands of students. Based on this carefully selected stratified sample, normative scores are developed.

Each of the derived scores that emerge from the standardization process, including percentile ranks, grade equivalents, and normal curve equivalents (NCEs), has a predefined relationship to the characteristics of the norm group. Thus, at the time the test is normed, 50% of the examinees will exceed the 50th percentile and the same percentage will fall below the 50th percentile. Derived score tables for the test battery are produced, and all scoring of student tests is referenced to these tables until the battery is either revised or, in rare instances, when it is renormed with no change in the content of the test.

Data from national probability samples are not usually collected for a test more often than every 5 to 8 years because it is impractical and economically infeasible to do. It would not be reasonable to ask or expect schools to administer tests to large numbers of students every school year in order to develop yearly norms based on a national probability sample. The cost of such testing would have to be passed on by the publisher to the schools and would add substantially to the cost of school testing programs.

In summary, most large test publishers follow the common and decades-old industry practice of revising and standardizing their achievement tests about every 8 years. The content is updated to reflect current curricula and instructional practices, and new norms are developed so that the test reflects levels of achievement that prevail during the school year in which the test is standardized. The dates of standardization are given wide publicity, and all purchasers of the test are aware of these dates.

Proper Interpretations of National Norms

Because norm-referenced tests are not normed yearly on a national probability sample, changes in national achievement between the norming years will be reflected in the norm scores for groups of students. For example, if national achievement levels decrease between normings, as they did from the late 1960s to the mid-1970s, students' norm-referenced scores will decrease, and more students will fall below the median (50th percentile) score established when the test was normed. On the other hand, when national achievement levels increase between normings, more students will exceed the median established when the test was originally normed. Regardless of the direction of national achievement trends, when a test is renormed, exactly half of the students will fall above and half will fall below the newly established median.

At this time, national achievement indicators all point to the fact that student achievement is generally on the increase. This increase is documented by the National Assessment of Educational Progress (NAEP); the Scholastic Aptitude Test (SAT) results; two Congressional Budget Office reports (1986, 1987); and data collected during recent test normings by CTB/McGraw-Hill (1985, 1987, 1988).

Thus, during a time of increasing national achievement, the students' normed test scores will rise between norming periods. More students will score above the median score established during norming than will fall below it. This confirms the sensitivity of the test norms to changes in achievement, one of the tests' primary functions. These normed test scores are valid measures of student growth. Although the reference year for the scores will be prior to the year in which the test scores are reported, the test scores provide accurate program and student information. The fact that the norm scores themselves refer to norming that took place during an earlier year in no way compromises the major purposes for administering an NRT or the usefulness of the scores for program evaluation, student instructional planning, or the monitoring of longitudinal trends. When interpreting the scores, the test user must simply be aware of the year that the tests were normed and the general direction of national achievement trends. Interpretive guidelines are found in relevant test-related materials produced by most publishers.

The Friends for Education report has received attention primarily as a result of its improper interpretations of score distributions for

norm-referenced tests between renorming years. The sensational, and apparently illogical, phenomenon of having too many students above the national average is the basis for the criticism leveled at the testing community by the report. This is a point that should be elaborated upon, because it may be misunderstood by others as well.

A naive interpretation of what an average (mean) represents is that half of the scores in a distribution will fall above and half will fall below the average. Although this is a common interpretation, it is not statistically correct. The report suffers from this misunderstanding, as illustrated by the following quote: "Standard principles of mathematics make it difficult for more than one half of any group to be above average" (Cannell, 1987). There is no mathematical principle that would cause this to be so. Depending upon the shape of the distribution of scores and the measure of central tendency that is selected to describe the scores, more or fewer than half the scores may be above or below the measure of central tendency. For example, the mean, or arithmetic average, does not necessarily split a distribution of scores into equal halves. An average that splits the distribution evenly will occur only in a symmetrical distribution. If the distribution is skewed, there may be many more scores above or below the average depending upon whether the distribution is negatively or positively skewed. The median (the 50th percentile), on the other hand, does separate a score distribution into equal halves. Thus, there is no a priori reason to believe that norm-referenced scores *should* separate the examinees into two equal halves, particularly during times of changes in national achievement trends.

Extended Extrapolations

The time-bound nature of normative interpretations is relatively straightforward to describe and understand. What becomes more difficult to evaluate are the social and educational implications that might be drawn from acknowledging that actual score distributions may differ increasingly from the published norms as a result of changes in achievement over time.

One way to determine the amount of change in achievement over time might be to survey states and school districts and, based on the aggregation of scores, determine the number of states and districts reporting above "average" (50th percentile) scores. Additionally, it might be possible to determine the proportion of students above the 50th percentile and the average national student score. Finally, to illustrate the rapidity with which standardization score distributions change, data could be collected the first year after norming and then an average of state and district scores could be calculated.

This task would be very difficult to do correctly. Different states and districts use different tests that are not on a common scale. The scores from all states and districts would have to be collected, placed on a common scale, and analyzed appropriately. There is no evidence that this has ever been done correctly.

This brute-force approach need not be the only mechanism to determine achievement trends over time, nor is it the best way. Achievement changes between normings are documented by the major publishers, and this information could be directly examined.

A third approach intended to monitor national achievement trends might be NAEP. But NAEP is also an imperfect panacea for determining achievement growth. There will always be quality-control issues, as evidenced by questions about recent NAEP survey results. NAEP is a valuable indicator of achievement trends, but like any method it is not absolutely perfect.

The fact is that various sources of information must be synthesized so that a complete picture of national trends can be obtained. Each type of assessment, via NRTs, CRTs, NAEP, or others, attempts to answer different questions in different ways. Each is valuable in providing a piece of the picture on the status of student learning. It is when we learn how to make artful syntheses that all of us will be closer to determining the status of achievement in America's schools.

It is unfortunate that during a time when national achievement trends are moving upward some might use that fact to suggest that one of the reasons for the upward movement is inadequate norming by test publishers and inappropriate teaching of test content by users for self-serving purposes. These are serious charges that should not be made without supporting evidence.

It must be stated that there is no logical reason *why* test publishers would wish to engage in inadequate norming. Test publishers have every incentive to make sure that their tests are completely objective and are administered properly and that their integrity as valid measures of performance stands unimpeached. Without such quality, test publishers would quickly find themselves with no customers.

Conclusions

To be sure, some of the concerns raised by Dr. Cannell are shared by all in the educational community. The time-bound nature of norms may not be well understood by some school personnel and the public. There may be abuses of tests and breaches of security. Some teachers and administrators may indeed disclose too much test content to the students. But the overwhelming majority of the educational community is doing its very best to administer tests and report test scores in a responsible fashion.

At least two examples of this come prominently to mind. The first is the way in which test publishers equate alternate forms within the same test battery over time. Thus, CAT Forms C and D (1977) were equated to CAT E and F (1985). Similarly, equating is done between different test batteries developed by the same test publisher, as was the case for CTBS Forms U and V (1981) and CAT E and F (1985). These equatings allow the test user to move from one version of a test to another and preserve longitudinal comparisons. The recent trend that has been observed in these equatings is that the derived scores from the most recently normed test are lower than for the earlier normed test. This is predictable in times of increasing national performance. The opposite would be true if national achievement trends were on the decrease. Explanatory material that helps the practitioner understand this phenomenon is

34-661 - 90 - 2

always provided.

The second example relates to the Annual National Normative Trend Data (NTD) published by CTB/McGraw-Hill. Research on this project began in 1984, when an emerging customer need was identified by the company. Customer comments about the desirability of obtaining more recent normative data were noted in market research efforts. Such data could be used to amplify the standardization norms and provide a more complete picture on the progress local school districts were making in their instructional efforts. After 3 years of research, the NTD service was offered to CTB customers. Score reports have been made available on an annual basis, for the standardization year as well as for the most recent norming. This service is a response to those educators who have been concerned about the time-bound nature of norm-referenced scores.

The test companies do their best, through many vehicles, to assist the test consumer in being a responsible user of test results. Indeed, reasonable testing programs, effectively implemented, are one of the reasons that achievement is increasing and that we are not currently in the decline phase that manifested itself in the late 1960s to the mid-1970s.

The assertion that scores are on the increase does have merit. Perhaps the positive side of this phenomenon should be stressed more. States and local school districts have committed considerable resources to improving the achievement levels of their students. All indicators of student achievement appear to converge on this fact, particularly for the elementary grades. The American public should be gratified that achievement is increasing.

Cannell (1987) charges that "inaccurate initial norms and teaching the test," rather than improved achievement, are reasons for improving scores on nationally normed tests. The problem with these allegations is that there is little, if any, evidence to support them. To the contrary, the body of independent evidence suggests that test norms provide a valid and useful reference in both the norming year and in subsequent years and that achievement at the elementary level has been increasing. If indeed there exist instances of abuse of test norms and of misunderstanding of their meaning by educators or the public in general, then the proper remedy should be to correct those instances rather than to make rash allegations about the adequacy of test norms or questionable teaching by educators.

References

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average.* Daniels, WV: Friends for Education.

Congressional Budget Office. (1986). *Trends in educational achievement.* Washington, DC: Author.

Congressional Budget Office. (1987). *Educational Achievement: Explanations and implications of recent trends.* Washington, DC: Author.

CTB/McGraw-Hill. (1977). *California Achievement Test, Forms C and D.* Monterey, CA: Author.

CTB/McGraw-Hill. (1981). *Comprehensive Test of Basic Skills, Forms U and V.* Monterey, CA: Author.

CTB/McGraw-Hill. (1985). *California Achievement Test, Forms E and F.* Monterey, CA: Author.

CTB/McGraw-Hill. (1987). *Annual national normative trend data.* Monterey, CA: Author.

CTB/McGraw-Hill. (1988). *Annual national normative trend data.* Monterey, CA: Author.

SRA Response to Cannell's Article

Audrey L. Qualls-Payne
Science Research Associates

The author defends SRA's norms, discusses some of the difficulties in pursuing Dr. Cannell's proposals, and points out that we need to monitor not just student achievement levels but also trends in curriculum.

Science Research Associates (SRA) recognizes the concerns expressed in John Cannell's article, "Nationally Normed Achievement Testing in America's Public Schools: How All 50 States Are Above the National Average." We differ, however, in our assessment of the situation and the proposed alternatives. According to the article, most schools in the nation perform at or above average on commercially available tests. This finding, as noted by Dr. Cannell, is not consistent with statistical theory, which says that half the students should be above and half below. Dr. Cannell expresses the opinion that this inconsistent statistical phenomenon results from using older tests, older norms, teaching to the test, statistical manipulation of the data by publishers, excluding special

Audrey L. Qualls-Payne, Psychometrician, Science Research Associates, Inc., 155 N. Wacker Dr., Chicago, IL 60606-1780. She specializes in achievement testing.

education students from the calculation of group averages, and inaccurate norms. He goes on to suggest that these problems can be eliminated by the use of one achievement test in all schools across the country with the concurrent development of annual norms. Our purpose is to examine Dr. Cannell's conclusions and offer alternatives to some of the issues raised in his report.

SRA's national norms are reliable and accurate indicators of national student performance at the time of standardization. The charge of statistical manipulation of data appears to result from Dr. Cannell's apparent misunderstanding of the purpose of the various types of test scores and subgroup norms. Schools may wish to compare their students' performance with, in addition to that of the national group, that of groups more similar in structure and student composition. For example, a nonpublic school may want to compare their students' performance with that of students from other nonpublic schools. The various test scores, in addition to status scores (i.e., percentiles and stanines), are offered to meet the many needs of our customers. Normal curve equivalents (NCEs) are required for Chapter 1 program evaluation. To assess longitudinal growth and determine functional levels, developmental scores, for example, standard scores and grade equivalents, are needed.

Dr. Cannell's alternative to the various standardized achievement tests is a national achievement test, which would require at least two major actions. First, this national achievement test would have to be normed annually with a representative group of students to have yearly norms. Second, new test forms would be needed for each administration to eliminate possible problems of teaching to the test and test security.

A project of this magnitude and complexity would be very difficult logistically and very costly. Two major logistic problems would be (a) obtaining curricular consensus on the test content and (b) obtaining or mandating national participation.

If yearly new forms are not an option but annual norming is, and if there truly is a substantial amount of teaching to the test, the problems noted in Dr. Cannell's analysis may not go away. If new forms of achievement tests are developed each year, thereby increasing test security, the need for annual norms diminishes significantly. Based on Dr. Cannell's analysis from schools with tight test security and literature on student growth, drastic shifts in student performance from one year to the next are rare. From a psychometric point of view, new norms are needed only when there is a *significant* shift in school curriculum and/or student performance.

As opposed to developing and standardizing new forms each year, a mechanism is needed to monitor changes in school curriculum and student performance. Whenever there is a change in either curriculum emphasis or achievement levels, new test forms should be developed and standardized. If the change is strictly a shift in student achievement, renorming is required. As a publisher, we must base our decision on when to issue new forms/new norms on a systematic monitoring system.

There are several ways to monitor student progress. One way to accurately spot when significant changes are taking place is to track student achievement on a regular basis (i.e., annually). The entire user group could be used for this purpose. The monitoring process should be capable of producing user-based norms, which can then be made available to all customers as an optional service in addition to the national norms.

There is at least one major problem with the user-based monitoring system. If the user sample is biased and unrepresentative of the national student population, significant changes noted in the user sample may not truly reflect changes at the national level. One way to resolve this problem would be to select a subset of schools from the user group and use it to monitor changes in curriculum and student achievement annually. The selected schools should be representative of the national population of schools with respect to geographic region and racial/ethnic and socioeconomic status. Once a set of schools is selected for this purpose, students in these schools can be tested on an annual basis and norms can be developed. As in the previous method, annual norms will be made available to customers as an optional service. Because of the representativeness of the schools selected for monitoring, one can, with a high degree of confidence, generalize results from this set of schools to the U.S. population of schools.

Because SRA recognized the value of a monitoring system, we are already in the developmental stages of implementing such a program.

Chairman HAWKINS. Well, thank you, Doctor.

The next witness is Dr. Walter Faithorn, Jr., retired business executive and volunteer teacher at the University of the District of Columbia.

Mr. FAITHORN. Yes, sir. Thank you, on behalf of Friends for Education which is the organization that I'm speaking for today. I'm not a doctor, Congressman Hawkins. I'm hopelessly outclassed in terms of professional lingo here today.

I'm from the Chicago manufacturing sector and my experience with the problem of the education of our children most recently in my career comes from the difficulty we've had in hiring people for our factories in Chicago who can read and write, and I mean people who have high school diplomas.

We went through many years of endeavoring to recruit factory employees in Chicago from those sections of town that had extremely high unemployment and poor economic conditions. We found lots of people who were willing to work and, as I said a moment ago, had high school diplomas—many of them—but they couldn't read the buses and the el-trains and the subways in order to get from where they live in south side of Chicago up to the near north side.

We started classes at my company to teach people how to read. But I didn't expect that was going to lead me to this room today. I've been retired for a few years, and I'm a volunteer teacher of English at the University of the District of Columbia.

I was asked by John Cannell, who is the President of Friends for Education, if I would substitute for him today. I'm sorry that he isn't here. I'm sorry for your sakes, as well as for my own. So I prepared a written statement which I submitted in many copies yesterday, and I was told, by the way, by one of your staff that I shouldn't read that today, I should be much more informal.

So let me begin by just quickly reviewing how the Friends for Education organization got started. John Cannell, who's just a youngster in my opinion—he's in his middle forties—and a kid I helped raise when he was just a youngster, is a physician—an M.D.—and was practicing in West Virginia. He opened a clinic up in the mountain woods territory for people who had never been able to get to doctors before. He lived in Beckly, West Virginia where his kids were going to public school.

He observed that they didn't seem to know from shynola what was going on in the world and they weren't learning anything. They didn't know whether France was the capital of Paris or Paris was the capital of France and all sorts of horror stories like that that I'm sure you've heard and read about from lots of other sources.

But at any rate, he went to see the school principal to talk about it. The principal and his staff were stunned that Dr. Cannell was unhappy because, as they pointed out to him, their school tested way above the National average. He was upset by this. The implications being that the National average was so terrible that he started writing school boards in Kentucky and West Virginia and then expanded his effort to include all 50 states over a period of several months, and found that virtually every school was reporting their schools and the students being above the National aver-

age. So he decided it was really a big fraud and that's what started his organization.

As I pointed out in what I submitted to you, he brought all this to the attention of the Department of Education and talked with the then Secretary Bennett who was quite skeptical of his findings at that time and authorized an investigation within the Department as to whether or not Cannell's claims were true, and found that they were. And lots of other sources have corroborated what Friends for Education discovered about these so-called national averages.

Cannell went on to pursue the matter and decided that not only were these norm-referenced tests misleading in terms of where students and schools stood, but that there was actually a sort of a conspiracy between the publishers and the school authorities, particularly superintendents, to continue these kinds of tests which made everybody so happy—the parents and everyone—because it made the schools look so good. Then the kids who were all testing above the National average when they came to take SATs or the Army tests, just were doing miserably and there was understandably a lot of confusion that resulted from that.

Well, all of these are well-known facts, I am sure, particularly among the professionals who are here today, but I thought it worth mentioning because that's what has made our organization as rank amateurs so upset with the present scene.

We believe very much in the points that were made by Dr. Haney. Nothing that he said about the need for much better testing—all of that is what Friends for Education stand for and are endeavoring to help expedite.

We also are very impressed with the work that's being done by the National Assessment of Educational Progress, which is within the Department of Education but has an outside governing board which seems to be giving it excellent impetus toward progressing this question of more intelligent testing.

We feel that the publishers who Dr. Faldet represents are extremely good businessmen. They've developed, as typical of good businessmen—and I can speak from personal experience on this point—a cozy relationship with their customers who are the school boards and superintendents, and they've promoted their product extremely effectively to the point where there are some 50 million tests a year of this sort given and the burden on children, as Dr. Haney has pointed out, is simply enormous. We'd like to see that greatly reduced.

Really, I don't have anything much more to say that will be of much value to your committee, but I'll be glad to answer what questions I can.

[The prepared statement of Walter E. Faithorn, Jr., follows:]

Friends for Education, Inc.
600 Girard, N.E.
Albuquerque, NM 87106
(505) 260-1745

5 June 1990

Walter E. Faithorn, Jr.
3800 Underwood Street
Chevy Chase, MD 20815

Augustus F. Hawkins, Chairman (California)
Committee on Education and Labor
U.S. House of Representatives
Washington, DC 20515

My dear Congressman Hawkins,

Thank you for inviting us to testify at the hearing on "Testing/Assessment/Evaluation, etc." of the Subcommittee on Elementary, Secondary and Vocational Education on 7 June 1990.

The views of Friends for Education on the subject of testing are, I believe, pretty well known. That was my inference, at any rate, from the comments of your administrative assistant, Dr. Dandridge, in a brief telephone exchange we had about arrangements for this meeting. I will summarize our position as follows:

We believe that well-conceived, properly designed and securely administered tests of students at a few crucial levels during their elementary and upper school years is absolutely essential in order for all to know what our children are learning, whether it is as much as we believe they should learn, and what schools are doing a superior, an average, or a poor job of teaching. We believe this kind of testing is necessary to the development and improvement of curriculum -- that good testing drives good curriculum.

We are sorely disenchanted with what is going on by way of testing, today, particularly the prolific use of commercially prepared and distributed "standardized achievement tests."

In 1988 we reported to educators, generally, and to William Bennett, then the Secretary of the U.S. Department of Education, in particular that the results of these tests were routinely compared to the scores of a "norm group" previously tested by the commercial publishers. The "norm group" had taken the test cold. Current scores of students now practiced in the same test are, quite naturally, higher, so Johnnie comes home with a computer printout that tells his parents that he is testing "above the national average," that by implication his school, his teachers,

his principal, the district superintendent, even the school board itself, are all "above the national average." Everybody is very happy. And even better, everybody in every other school district in all fifty states is equally happy because every kid in every school is testing "above the national average." We contacted more than 3,500 school districts in all fifty states and what they reported to us said, in effect, that 70% of our school children were testing above the publisher's "national norm" on commercial norm-referenced achievement tests.

Secretary Bennett called this the "Lake Woebegone" effect; he was at first skeptical of our report; he authorized an $80,000 study by his department to check out our story; his study confirmed our findings.

Because it became immediately apparent that the widespread use of these commercial, norm-referenced, achievement tests so profoundly affect the feelings of the public at large toward their schools and all the personnel connected with them, these tests became known as "high stakes" tests. When scores are high, everyone's job is secure; if scores are low, heads might roll. It also did not take a high order of perception on our part to realize that "high stakes" testing requires high security in handling them. Many and all manner of unsolicited letters began to arrive, mostly from teachers, about how they were required to "teach the test," to cheat by giving out answers before the tests, by changing answers after the test, by keeping predictably poor pupils from taking the test (invariably children from minorities and/or other disadvantaged groups), all such manipulations to better insure better results to be better than the "national average."

We investigated these allegations, at least as many as we could, thanks to a small grant from the Kettering Foundation, and to our satisfaction found them generally true -- if anything, understated -- just the proverbial tip of the iceberg. We think all kinds of cheating is going on in respect to these tests and we think the big, commercial publishers of these tests know it and at best look the other way. I do not use the term, "bureaucracy," pejoratively. Our whole society is a bureaucracy -- big business, government, professions, nonprofit organizations, public education; none of it could function without bureaucracy. And the public school bureaucracy quite naturally takes great comfort in being able to report "above national average" test results.

The principal victims of this scam are, of course, the children, and the most vulnerable of these victims are minority children and those otherwise disadvantaged by poverty or other calamity. And when Johnnie does really poorly on college entrance and SAT exams and on Armed Services Vocational Aptitude Battery scores -- when the emperor comes by without any clothes on -- we all look at each other in puzzlement, and parents wring their hands in confusion.

We think we should all be wringing our hands -- but more because of imminent peril than confusion. Our kids compare miserably with their opposite numbers in the world's other leading industrial nations. Just to name a couple, how far behind our two former mortal enemies is all this going to leave us in a few short years?

The U.S. Department of Education will not carry its investigation of our work beyond confirming, as it did, our "Lake Woebegone" exposé of performance "above the national average." In an effort to better prepare myself for this hearing, I met with a group of senior officials in the office of the Assistant Secretary for Educational Research and Improvement, and I was told that our allegations of cheating, fraud, and deceit were of an "anecdotal" nature and did not lend themselves to rigorous and objective verification. These gentlemen went on to say -- and to me this was the more important part of their answer -- that not the Congress, nor the States, nor the local school boards (all 15,000 plus of them) want the U.S. Department of Education messing around in matters of this sort -- telling them what they are doing wrong, how this State compares to that State, or this school district compares to that, etc. They didn't talk about money, but I suspect that it is also quite acceptable to the White House Chief of Staff, among others, that the Department of Education not be making plans or noises about things that cost serious money. We really wonder if President Bush believes that the federal contribution of 6¢ for every dollar spent on public education is enough.

Nonetheless, we are impressed by the work being done by the National Assessment of Educational Progress (NAEP) and its governing board, the National Assessment Governing Board (NAGB) towards setting achievement levels defining what students ought to know at different grades. Because it seems to us that our kids are operating under an unreasonably onerous load of testing in their schools, today, we applaud the NAGB approach of limiting the testing they propose to the fourth, eighth and twelfth grades. We very much approve of the way in which they are going about developing a broad national consensus for defining standards to be used in accomplishing improvement. We are concerned, however, about a possible reduction in the rigor with which test security will be practiced.

As many of you may know, Friends for Education is a small group of rank amateurs founded and energized by a young physician, John J. Cannell, in general practice in the back woods of West Virginia where he didn't think his kids were learning very much in the public school. When he talked to the principal about his concern, he was told that he should be happy, just as all the other parents were, because his school tested "way above the national average." That's how it all stared. A few parents joined him, and because he could not afford the time or money to come to this hearing, I was pressed into service (as an old, old

friend who helped him out a bit when he was a youngster). He is now living in New Mexico, is on the faculty of that University's medical school, and is also carrying on demanding post-graduate studies.

I am even a ranker amateur, a retired business executive from the manufacturing sector in Chicago, now a volunteer teacher at the University of the District of Columbia, and a volunteer representative, today, for Friends for Education. (Permit me please, to note, parenthetically, that the amount of work I have had to do in order to do any justice at all to this opportunity to meet and speak to this Committee has filled me with wonder at how Congress ever gets anybody to testify at hearings who is not well supported by staff, facilities, money and highly paid professional spokesmen.

I wish my classmate, Francis Keppel, were still on the scene. (One should not infer from that that we were close friends; actually we were mere acquaintances, but I admired him greatly.) I feel certain that there would be a much more aggressive attitude in the Department of Education today. Outrage and fury over the ridiculously heavy load of endless testing, and so much of it deceitful at that, would drive him, I think, to the doors of Congress for money and authority to do at least as much to protect children from fraud as a Department of Agriculture meat inspector does in his field.

I think he would invite the Congress's attention to the recent T.V. edition of CBS's 60 Minutes which documented widespread cheating by school authorities in South Carolina and how the blame and punishment was put on teachers. Francis might well say to you, "If you won't let my department do anything about this, if you won't let me protect the first line of victims, our children, and the second line, our teachers, then you do it. It won't be the first time Congress has been goaded into action by a television program. When you want to, Congress, you can do some pretty heavy-duty investigating."

Thank you very much.

Respectfully,

Walter E. Faithorn, Jr.

EDUCATION

Newsweek

We need to produce students who know how to think. And we need new tests to help us.

Since last September's national education summit in Charlottesville, Va., a committee of governors and White House staff members has been meeting in Washington to define the nation's educational goals. The panel hopes to flesh out its list this month, just in time for President Bush to include it in his State of the Union address to Congress. That's the easy part. Everyone favors reducing the dropout rate and encouraging local creativity. But for once, everyone also recognizes that platitudes won't be adequate to the task. The hard part will come later, when the committee is expected to propose a national yardstick that will enable Americans to know how their students and schools are faring.

"The big question," says committee co-chairman Carroll Campbell, the governor of South Carolina, "is how do you measure success—against what test?"

For educators, particularly those bent on reform, that is the question for the 1990s. It is, of course, no small irony that a system devoted to testing any child old enough to sit still now stands accused of gross failure of tests. Depending on the critic, the nation's vaunted standardized exams: (a) measure the wrong skills; (b) distort classroom instruction; (c) falsely reassure parents; (d) discriminate against the underprivileged. The push for better tests comes from several sources. Elected officials are demanding more from the education mandarins and need a way to hold them accountable. Parents, from Miami to Chicago, are being invited to share in the management of neighborhood schools; they, too, are unwilling to rely on the tyranny of anecdote or true-false quizzes. And employers say they no longer want people who have mastered just the basics—although some days they wish they had more of them—but they need people who can think. In short, the consumer movement has finally entered the schoolhouse.

Newspapers rank schools and districts by their test scores, real-estate agents use test scores to identify the best schools as selling points for expensive housing, school superintendents can be fired for low scores, and teachers can receive merit pay for high scores," says Lorrie Shepard, an education professor at the University of Colorado.

Forty-two states give mandatory achievement tests. Some, like New York and California, develop their own tests. Others use one of four leading commercial brands. In addition, many local districts and even individual schools require standardized exams on their own. And the federal government's National Assessment of Educational Progress (NAEP) regularly tests a small random sample of fourth, eighth and 12th-grade students nationwide. A recent study by FairTest, a Boston-based advocacy group, found that U.S. public schools administered 105 million standardized tests to 39.6 million students in the 1986-87 school year alone—an average of more than 2.5 standardized tests per student per year. But for all the many choices, says Ernest Boyer, president of The Carnegie Foundation for the Advancement of Teaching, "Most of our current efforts at assessment have been woefully inadequate, fragmented and even destructive."

Consider the "Lake Wobegon" effect, named after Garrison Keillor's fictional town where "the women are strong, the men are good looking and all the children are above average." After years of treating troubled West Virginian adolescents whose problems invariably included poor academic performance, psychiatrist John Jacob Cannell wondered why the local schools claimed that his patients had high scores on their nationally normed standardized tests. Cannell requested test scores from all 50 states. He discovered that each one claimed to be "above the national average" or the statistical norm. Cannell has filed a complaint with the Federal Trade Commission charging the four leading commercial test publishers with deceptive trade practices.

What explains Lake Wobegon? Test publishers only update their norms when they create a new edition of a test, usually every seven years. In the meantime, students are compared to a national standard that is sometimes more than half a decade out of date. If achievement is improving nationally, the use of old norms will make a district appear to be doing better than it would be if current norms. The better scores may not be due to better learning; some savvy schools now give classes in testing tactics. And some teachers and principals now engage in outright cheating. For example, last spring, a Greenville, S.C., high school teacher discovered a student cheating on a test. Teacher Nancy Yeargin, considered by colleagues to be one of the best teachers in the school, admitted that she had distributed the questions and answers two days before a standardized exam to two low-ability geography classes. She was fired and prosecuted under a South Carolina law that makes breaching test security a crime. Yeargin avoided spending 90 days in jail by pleading guilty last September and paying a $500 fine. Clever administrators needn't go that far; instead some simply encourage slower pupils to be absent on the day of an exam.

A team of education professors from UCLA and the University of Colorado recently completed a more scientific version of Cannell's study that compared student test scores from various states. They found Cannell is essentially correct. But they don't have a simple solution. "The problem is that too much is being expected of these tests," says Robert Linn, a professor of education at the University of Colorado.

The growing dissatisfaction with most standardized tests has led to a search for alternatives. For large-scale assessment, some experts favor expanding the federally funded NAEP. Over its 20-year history, NAEP has earned an enviable reputation among educators. It's a test that combines multiple-choice, essay and problem-solving questions. It's given on a random basis, to only a few kids in any classroom; teachers, then, can't "teach the test." And its norms are developed so carefully that they're trusted. However, until recently federal law prohibited the use of NAEP scores to compare states or districts. Congress finally permitted an experimental state-by-state comparison of scores for the eighth-grade math test in 1990. So far 37 states have agreed to participate.

The NAEP board would like to expand the comparisons across the board but some educators fear that will inevitably ruin the test. "If NAEP scores become so important to educators and policymakers that classroom instruction is tailored to it, its value as an indicator of achievement will be threatened," says Daniel Koretz, an analyst at the Rand Corporation. NAEP board chairman Chester E. Finn thinks the nature of the test procedure will prevent that from happening. "We've got a very good instrument and we've got an enormous appetite for information on how much kids are learning," he says.

Claudia Kalb and Pat Wingert

The New York Times

THE NEW YORK TIMES EDUCATION WEDNESDAY, FEBRUARY 17, 1988

Standardized Test Scores: Voodoo Statistics?

Schools suffering from 'Lake Wobegon effect'

By Edward B. Fiske

Garrison Keillor, the folk humorist, has provoked many a smile by his description of the mythical town of Lake Wobegon as a place where the women are strong, the men good-looking and "all the children are above average."

But looking closely at the results of the estimated 50 million standardized achievement tests taken by American schoolchildren every year, it seems that such fantasies are no longer a laughing matter.

For several years virtually every state education department — and even the most urbanized local school districts — have released standardized test scores showing that their children are reading, writing and calculating above the national average. Since this by definition is impossible, test makers and educators have been accused of playing statistical or educational shell games.

Last week Chester E. Finn Jr., the assistant U.S. secretary of education in charge of research, recalled both sides into his office to explore the issue. He concluded that the standardized test scores used to evaluate public schools were not always what they appeared to be. "Maybe it's time for the Department of Education to do something by way of providing some information to consumers," he said afterward.

What Finn described as the "Lake Wobegon effect" was first raised in a systematic way by Dr. John J. Cannell, a family physician in Beaver, W.Va., who was concerned about the problems of low self-esteem and depression he saw in many of his teenage patients. "I noticed a discrepancy between their academic performance and the grade level to which they were assigned," he said.

When Cannell heard a report one day from his state's Education Department that schoolchildren in West Virginia, which has one of the highest illiteracy rates in the nation, were performing above the national average, his concern turned into anger.

He formed a non-profit organization, Friends for Education Inc., and canvassed state education departments around the country. "We could not find any state that was below the national average," he said.

Participants in last week's meeting, called by Finn to explore Cannell's charges, said what struck them the most was that none of the two dozen people present — not even the test makers — took issue with his findings.

"There's no dispute that test scores are rising," said David G. Deffley, general manager of CTB-McGraw-Hill, publisher of the California Achievement Test and other tests. "The dispute comes about why."

At least these four reasons are usually suggested:

▲ Definitions of "average" are out of date. The major tests are first given to a scientific sample of students around the country. Their scores become the benchmarks, or norms, for determining whether those who follow are scored above, at or below the national average. But norms for many of these tests have not been reset for six or seven years. Testers say that schools have been getting better and the average has been rising. Consequently many students who might otherwise now be scored "below average" are still "above average" compared with the early 1980s sample groups.

▲ Schools pick tests that match their curriculums. This means that their students, unlike many of those in the sample, will find a close fit between the questions and what they have been taught. "There's an upward bias," said Deffley. "By virtue of the match, you're likely to have higher scores."

▲ Once teachers become familiar with the tests, they tend to alter their teaching to anticipate what their students will encounter.

▲ There are no industry standards on what students take the test. For the trial tests, many districts give the exams to all students, including those with learning problems. But many of the districts using the test exclude the scores of such students. "I find that reprehensible," said Finn.

Some explanations are themselves disputed, beginning with the assumption that schools have improved in the 1980s. Although primary school scores have risen, high school scores have not. The Educational Records Bureau, which specializes in testing students in private schools and in wealthy suburban districts, reports that scores at all grade levels have been "stable since 1979-80."

There is also debate over how far schools go to align their curriculums with the tests. Cannell charges that many schools are giving their students actual test items. Test makers acknowledge that some schools use the same form of the tests every year, but they argue that the outright "cheating" alleged by Cannell is not widespread.

When all is said and done, everyone seems to agree that the standardized testing in this country is structured so that except in rare periods when student learning is on the decline, it is impossible to have a test where half of the students will be reported to be above average and half below.

"The testing industry wants to sell lots of tests, and the school superintendents desperately need high and improving scores," said Cannell. "Nobody is disappointed."

Is there a moral or ethical issue here? The fact that educators and test makers are making public statements that they know are misleading to parents and taxpayers would suggest that there is. Cannell, however, says that the issue goes beyond old-fashioned truth to one of justice.

"The appearance of high scores allows school districts to continue turning out functionally illiterate children," said Cannell.

Finn, who acknowledges he did not realize the full extent of the testing paradox, suggested that it would be a "fine idea" to hold any future meetings in the Chatterbox Cafe.

That's the place in Lake Wobegon that serves up Powdermilk Biscuits wholesome enough to "give shy persons the strength to get up and do what needs to be done."

Fiske writes for The New York Times.

The Boston Globe

MONDAY, NOVEMBER 20, 1989

Critics press for alternatives to standardized tests

By Muriel Cohen
GLOBE STAFF

Complaints about the abuse of student testing procedures and scores are prompting educators and test publishers to take steps to reduce cheating and to stop the use of test results as measures of school quality.

Test-related problems have become so pervasive, and have generated such anger, controversy and confusion, that a national clearinghouse to find better ways to test and report results is being established at the University of North Carolina. Albert Shanker, president of the American Federation of Teachers, recently called for an immediate stop to standardized tests at the elementary and secondary school levels.

An increase in calls for accountability have led to the use of tests to judge the effectiveness of teachers and administrators. Scores on state achievement tests and national tests such as the Scholastic Aptitude Test are used as a measure of high school or school district quality.

For example, one condition for the renewal of Boston schools Superintendent Laval Wilson's contract is evidence that test scores have risen under his tenure. Boston is using tests that were "normed" six years ago, according to Mary Ellen Donahue of the system's testing office.

"Norming" is a process used by publishers to determine what an average score would be for a given grade level.

Testing companies do not establish a norm every year because they say it would be too expensive; the "norming" process is done every seven or eight years, which gives teachers familiarity with the tests and enables them to "teach to the test," critics say.

Researcher sees wide cheating

Pressure to produce ever-higher scores in reading, writing and math has led to a good deal of cheating among school teachers, principals and superintendents, according to Dr. John Jacob Cannell, a resident psychiatrist at the University of New Mexico.

The test publishers are beginning to respond to educators' complaints. McGraw-Hill, which produces the widely used California Basic Skills tests, is tightening up security and trying to find a shortcut for establishing norms so there will be a new average each year.

"Industry standard for norming is seven to eight years," said John Stewart, of McGraw-Hill's testing unit, based in Monterey, Calif.

"We now print on all score reports the date of the norming so that schools can't hide behind old dates. We no longer put answer keys in test packages to schools, but send them to the district coordinator to help avoid cheating," Stewart said.

"We are looking into developing a new test on an annual basis. Now we can amortize the cost over an eight-year period, but if we have to produce one annually, it will send the price up."

It was Cannell who found that most school systems in the country claim that their students perform above national norms, which is mathematically impossible.

"Scores on local norm-referenced achievement tests have improved so dramatically that by early 1988 all 50 states were testing above the publisher's national norm," wrote Cannell in his newest book, "How Public Educators Cheat on Standardized Achievement Tests."

A previous Cannell report "documented how 70 percent of American elementary children, 90 percent of American school districts and all 50 states were testing above the publisher's national norm ... instead of the expected 50 percent."

Cannell said there is a general lack of security in the storing and distribution of tests. "Teachers and principals have time to examine the tests and coach their students with similar kinds of questions on comparable time schedules," he said.

"My battle is getting accountability in testing," Cannell said in a telephone interview. "Teachers will cheat to get scores up when the stakes are high. Superintendents and school administrators are the main beneficiaries of teacher cheating and many teachers don't think they are cheating."

"Maine does it right, California does a creditable job and Massachusetts doesn't do too badly when compared to others," Cannell said.

In a state-by-state analysis, he found that "the Maine Educational Assessment is reported with scaled scores and is no longer equated with nationally normed tests. Teachers may not obtain the test booklets until the day before testing, the booklets are shrink-wrapped when delivered for the schools and testing is monitored by state officials."

Maine education commissioner Eve Bither said that committees of Maine teachers developed the test over a two-year period. The assessment is given to all fourth, eighth and eleventh grade students.

"We are very careful with the tests. In a class of 12 in fourth grade science, for instance, it is possible that each child will have a different set of questions," Bither said.

Bither concurred with Cannell on the standardized tests.

"When the Maine Educational Assessment came out for the first time, we found that the nationally standardized tests painted a far rosier picture, that everybody here was above average, than what the MEA found," said Bither.

Walter Haney of the Boston College Test Center said that Cannell has done a real service by bringing parental attention to test results.

THE ARIZONA REPUBLIC

Educators aid cheating, report says

Seek job security by helping students on standardized tests

New York (AP) — Educators desperate to demonstrate excellence are helping students cheat on standardized tests, contributing to scores in 46 states that are misleadingly "above average," a report charges.

At the same time, test security in virtually all states remains "totally inadequate," according to "The 'Lake Wobegon Report': How Public Educators Cheat on Achievement Tests."

The cheating, which includes teachers and principals coaching students on test questions, giving students more than the allotted time to take tests, and even altering answer sheets, is contributing to inflated scores, the report said.

All but two states — Louisiana and Arizona — are reporting "above-average" or inflated scores, according to the 50-state survey. Those states recently started using new tests.

Eighty-three percent of 5,143 elementary school districts and 73 percent of 4,501 secondary school districts surveyed are reporting standardized achievement test scores above national norms, according to the report.

The study was conducted by Friends for Education, an educational watchdog group headed by an Albuquerque, N.M., physician, John Jacob Cannell. It was funded by a $25,000 grant from the Kettering Family Foundation in Dayton, Ohio.

Dr. Cannell made headlines in November 1987 with a report documenting that students were scoring "above average" on standardized tests in all 50 states at that time.

That report asserted that scores on such "norm-referenced" tests — designed so that only half those taking it should score above the 50th percentile — were artificially high largely because the norms were not being updated often enough by test publishers.

The resulting overly bright picture of student achievement became labeled the "Lake Wobegon Effect," after author Garrison Keillor's mythical Minnesota town where "all the children are above average."

Test publishers have responded that it's expensive to renorm tests as frequently as critics like Dr. Cannell demand. They say the improved scores, in fact, show that schools are getting better. And they defend their tests as useful tools to identify pupil or group strengths and weaknesses.

Those tests, taken by students in all 50 states, include the California Achievement Test, the Stanford Achievement Test, the Metropolitan Achievement Test, the Science Research Associates Test, the Comprehensive Test of Basic Skills, and the Iowa Test of Basic Skills.

Cannell's accusations were largely confirmed in 1986 by a follow-up study sponsored by the U.S. Education Department.

Many states have leaped on new standardized test requirements during the 1980s as part of the drive for higher school standards. Teachers, principals and school administrators have found themselves under pressure to make their schools, and themselves, look better through higher scores.

In some states, teachers' careers can now be made or broken and school districts can be subject to state takeover partly on the strength of standardized test scores.

Cheating Admitted

"Those tests that were once used only as instructional aids now assess class achievement, school achievement and district achievement through student's scores," the report said.

Dr. Cannell said Friday that in addition to the test data he gathered from all 50 states, he placed an ad in the trade journal "Education Week," inviting educators to describe cases of test cheating.

Dr. Cannell said he received more than 300 letters from present and former teachers and school administrators admitting that they or colleagues had tampered with tests or helped students improperly. All demanded anonymity, Dr. Cannell said.

One Tennessee teacher wrote that teachers in his school "spent the morning teaching the test and the afternoon giving it."

In a number of states, Dr. Cannell wrote, scores are "much higher than any other indicators that often correlate with school achievement."

Sixty-five percent of Georgia's second-graders, for example, scored above average on the Iowa Test of Basic Skills, and 75 percent of Kentucky's third-graders scored above national norms on the Comprehensive Test of Basic Skills, "despite the fact that Georgia and Kentucky have among the lowest literacy rates, lowest college entrance scores and lowest Armed Services Vocational Aptitude Battery scores in the nation."

Few states randomly audit test scores to uncover improprieties, according to the study. California, one of a few that does, caught 30 schools cheating during the past three years on its statewide achievement testing program.

According to the report:

— Only 16 states forbid test administrators to receive the tests earlier than the day the test is to be given. Four others are instituting such guidelines.

— Six states forbid teachers from reading the test booklets; seven others plan to.

— Only a dozen states require that test booklets be sealed.

Drafts of the report were reviewed this summer by a dozen testing authorities, child psychiatrists and educators, including Chester Finn, Jr., former assistant U.S. education secretary and now a professor of education at Vanderbilt University.

Finn called the report "a constructive and useful piece of work."

"If Cannell is right, and his track record is such that he probably is, states are so lax and sloppy in organizing test security that it's like letting Exxon monitor water quality in Prince William Sound."

One Tennessee teacher wrote that teachers in his school "spent the morning teaching the test and the afternoon giving it."

THE WALL STREET JOURNAL.

Western Edition THURSDAY, NOVEMBER 2, 1989 Denver, Colorado

Classroom Scandal

Cheaters in Schools May Not Be Students, But Their Teachers

Pressure to Bolster Test Scores Drove Beloved Instructor, Nancy Yeargin, to Crime

Is She a Martyr or Villain?

By Gary Putka
Staff Reporter of The Wall Street Journal

GREENVILLE, S.C. – Cathryn Rice could hardly believe her eyes. While giving the Comprehensive Test of Basic Skills to ninth graders at Greenville High School last March 16, she spotted a student looking at crib sheets.

She had seen cheating before, but these notes were uncanny. "A stockbroker is an example of a profession in *trade and finance*. . . . At the end of World War II, Germany surrendered before Japan. The Senate-House conference committee is used when a bill is passed by *the House and Senate in different forms*."

Virtually word for word, the notes matched questions and answers on the social studies section of the test the student was taking. In fact, the student had the answers to almost all of the 40 questions in that section. The student surrendered the notes, but not without a protest. "My teacher said it was OK for me to use the notes on the test," he said.

Nancy Yeargin

The teacher in question was Nancy Yeargin – considered by many students and parents to be one of the best at the school. Confronted, Mrs. Yeargin admitted she had given the questions and answers two days before the examination to two low ability geography classes. She had gone so far as to display the questions on an overhead projector and underline the answers.

Mrs. Yeargin was fired and prosecuted under an unusual South Carolina law that makes it a crime to breach test security. In September, she pleaded guilty and paid a $500 fine. Her alternative was 30 days in jail.

Her story is partly one of personal downfall. She was an unsinning teacher who was lauded and inspired students, but she will probably never teach again. In her wake she left the bitterness and anger of a principal who was her friend and now calls her a betrayer; of colleagues who say she brought them shame; of students and parents who defended her and insist she was treated harshly; and of school-district officials stunned that despite the baldfaced nature of her actions, she became something of a local martyr.

Possible Motivation

Mrs. Yeargin's case also casts some light on the dark side of school reform, where pressures on teachers are growing and where high-stakes testing has enhanced the temptation to cheat. The 1987 statute Mrs. Yeargin violated was designed to enforce provisions of South Carolina's school-improvement laws. Prosecutors alleged that she was trying to bolster students' scores to win a bonus under the state's 1984 Education Improvement Act. The bonus depended on her ability to produce higher student-test scores.

"There is incredible pressure on school systems and teachers to raise test scores," says Walt Haney, an education professor and testing specialist at Boston College. "So efforts to beat the tests are also on the rise." And most disturbing, it is educators, not students, who are blamed for much of the wrongdoing.

A 50-state study released in September by Friends for Education, an Albuquerque, N.M., school-research group, concluded that outright cheating by American educators is common. The group says standardized achievement test scores are greatly inflated because teachers often "teach the test," as Mrs. Yeargin did, although most are never caught.

Evidence of widespread cheating has surfaced in several states in the last year or so. California's education department suspects adult responsibility for erasures at 90 schools that changed wrong answers to right ones on a statewide test. After numerous occurrences of questionable teacher help to students, Texas is revising its security practices.

Advance Notice

And sales of test-coaching booklets for classroom instruction are booming. These materials, including Macmillan/McGraw-Hill School Publishing Co.'s Scoring High and Learning Materials – are nothing short of sophisticated crib sheets, according to some recent academic research. By using them, teachers – with administrative blessing – telegraph to students beforehand the precise areas on which a test will concentrate, and sometimes give away a few exact questions and answers. (See related article on page A14.) Use of Scoring High is widespread in South Carolina and common in Greenville County, Mrs. Yeargin's school district.

Experts say there isn't another state in the country where tests mean as much as they do in South Carolina.

Under the state's Education Improvement Act, low test scores can block students' promotions or force entire districts into wrenching, state-supervised "interventions" that can mean firings. High test scores, on the other hand, bring recognition and extra money – a new computer lab for a school, grants for special projects, a bonus for the superintendent.

Critics say South Carolina is paying a price by stressing improved test scores so much. Friends of Education rates South Carolina one of the worst seven states in its study on academic cheating. Says the organization's founder, John Cannell, prosecuting Mrs. Yeargin is "a way for administrators to protect themselves and look like they take cheating seriously, when in fact they don't take it seriously at all."

Paul Sandifer, director of testing for the South Carolina department of education, says Mr. Cannell's allegations of cheating "are purely without foundation," and based on unfair inferences. Partly because of worries about potential abuse, however, he says the state will begin keeping closer track of achievement-test preparation booklets next spring.

Students' Perspective

At Greenville High School, meanwhile, some students – especially on the cheerleading squad – were crushed. It's hard to explain to a 17-year-old why someone they like had to go, says Mrs. Ward Soon. T-shirts appeared in the corridors that carried the school's familiar red-and-white GHS logo on the front. On the back, the shirts read, "We have all the answers."

Many colleagues are angry at Mrs. Yeargin. She did a lot of harm, says Cathryn Rice, who had discovered the crib notes. "We work damn hard at what we do for damn little pay, and what she did cast unfair aspersions on all of us."

But several teachers also say the incident casts doubt on the wisdom of evaluating teachers or schools by using standardized test scores, says Gayle Key, a mathematics teacher. "The incentive pay thing has opened up a can of worms. There may be others doing what she did."

THE WALL STREET JOURNAL.

Western Edition THURSDAY, NOVEMBER 2, 1989 Dayton, Columbus

Tests Often Match Materials in Kits And Study Booklets

How a California Exam Has Same Question Included In Commercial Workbook

By Gary Putka
Staff Reporter of The Wall Street Journal

Since chalk first touched slate, schoolchildren have wanted to know: What's on the test? These days, students can often find the answer in test-coaching workbooks and worksheets their teachers give them in the weeks prior to taking standardized achievement tests.

The mathematics section of the widely used California Achievement Test asks fifth graders: "What is another name for the Roman numeral IX?" It also asks them to add two-sevenths and three-sevenths.

Worksheets in a test-practice kit called Learning Materials, sold to schools across the country by Macmillan/McGraw-Hill School Publishing Co., contain the same questions. In many other instances, there is almost no difference between the real test and Learning Materials. What's more, the test and Learning Materials are both produced by the same company, Macmillan/McGraw-Hill, a joint venture of McGraw-Hill Inc. and Macmillan's parent, Britain's Maxwell Communication Corp.

Parallels to Tests

Close parallels between tests and practice tests are common, some educators and researchers say. Test preparation booklets, software and worksheets are a booming publishing subindustry. But some practice products are so similar to the tests themselves that critics say they represent a form of school-sponsored cheating.

"If I took [these preparation booklets] into my classroom, I'd have a hard time justifying to my students and parents that it wasn't cheating," says John Kaminski, a Traverse City, Mich., teacher who has studied test coaching. He and other critics say such coaching aids can defeat the purpose of standardized tests, which is to gauge learning progress.

"It's as if France decided to give only French history questions to students in a European history class, and when everybody aces the test, they say their kids are good in European history," says John Cannell, an Albuquerque, N.M., psychiatrist and founder of an educational research organization, Friends for Education, which has studied standardized testing.

Standardized achievement tests are given about 10 million times a year across the country to students generally from kindergarten through eighth grade. The most widely used of these tests are Macmillan/McGraw's CAT and Comprehensive Test of Basic Skills; the Iowa Test of Basic Skills, by Houghton Mifflin Co.; and Harcourt Brace Jovanovich Inc.'s Metropolitan Achievement Test and Stanford Achievement Test.

Sales figures of the test prep materials aren't known, but their reach into schools is significant. In Arizona, California, Florida, Louisiana, Maryland, New Jersey, South Carolina and Texas, educators say they are common classroom tools.

Hrisk Sales

Macmillan/McGraw says well over 10 million of its Scoring High test preparation books have been sold since their introduction 10 years ago, with most sales in the last five years. About 20,000 sets of Learning Materials teachers' binders have also been sold in the past four years. The materials in each set reach about 30 students. Scoring High and Learning Materials are the bestselling preparation tests.

Michael Kean, director of marketing for CTB/Macmillan/McGraw, the Macmillan/McGraw division that publishes Learning Materials, says it isn't aimed at improving test scores. He also asserted that exact questions weren't replicated. When referred to the questions that matched, he said it was coincidental.

Mr. Kaminski, the schoolteacher, and William Mehrens, a Michigan State University education professor, concluded in a study last June that CAT test versions of Scoring High and Learning Materials shouldn't be used in the classroom because of their similarity to the actual test. They devised a 69-point scale—awarding one point for each subskill measured on the CAT test—to rate the closeness of test preparatives to the fifth grade CAT.

Because many of these subskills—the symmetry of geometrical figures, metric measurement of volume, or pie and bar graphs, for example—are only a small part of the total fifth grade curriculum, Mr. Kaminski says, the preparation kits wouldn't replicate too many, if their real intent was general instruction or even general familiarization with test procedures. But Learning Materials matched on 66.5 of 69 subskills. Scoring High matched on 64.5.

Fifth Grade Exam

In CAT sections where students' knowledge of two-letter consonant sounds is tested, the authors noted that Scoring High concentrated on the same sounds that the test does—to the exclusion of other sounds that fifth graders should know.

Learning Materials for the fifth grade contains at least a dozen examples of exact matches or close parallels to test items.

Rick Brownell, senior editor of Scoring High, says that Messrs. Kaminski and Mehrens are ignoring the need students have for becoming familiar with tests and testing format. He said authors of Scoring High "scrupulously avoid" replicating exact questions, but he doesn't deny that some items are similar.

When Scoring High first came out in 1979, it was a publication of Random House. McGraw-Hill was outraged. In a 1980 advisory to educators, McGraw-Hill said Scoring High shouldn't be used because it represented a "parallel form" of the CAT and CTBS tests. But in 1986, McGraw-Hill purchased the Random House unit that publishes Scoring High, which later became part of Macmillan/McGraw. Messrs. Brownell and Kean say they are unaware of any efforts by McGraw-Hill to modify or discontinue Scoring High.

Rocky Mountain News

Vincent Carroll

The great testing lie

A prediction: When grade-school students take achievement tests this spring, most of their scores will rise from a year ago.

Now the bad news: Those scores will be virtually meaningless.

From coast to coast, school districts conspire in The Great Testing Lie. They permit the same standardized tests to be administered year after year. Whether deliberately or not, teachers apparently adjust curriculums to emphasize test material. There is no other way to explain why so many districts consistently report rising scores.

Achievement tests for younger students especially generate laughably implausible results year after year, and no one seems to mind.

Well, almost no one. A doctor in West Virginia cared so much he spent $11,000 of his own money to compile and assess achievement test results from every state and hundreds of districts. So far, no one disputes John Jacob Cannell's remarkable conclusions, not even the testing companies themselves. Among his findings:

■ About 90 percent of U.S. school districts claim to be above average in student achievement, and most announce year-to-year improvement.

■ Every Southern state tests above the national average except Mississippi, and even it is at the mythical national "norm."

■ More than 70 percent of American children are told they test above the national average.

■ Twenty-six states test on a statewide basis, and all report above-average scores. Six others, which have developed their own tests and give them statewide, test above average, too.

The problem isn't merely that the same tests are reused. Most elementary students score better than average even on brand-new achievement tests, raising doubts about the accuracy of the norms themselves.

Georgia, for example, should have everything going against it in the testing derby — large numbers of disadvantaged children, high dropout rates, low college-entrance scores among high-school seniors — and yet Georgia's second-graders scored above the 60th percentile nationally in every category the first year a revised Iowa Test appeared.

Not every district reports above-average scores, of course. Achievement is so bad in some big cities that even artificially low norms only partly disguise the fact.

But give those districts time. If they stick with the same tests for the next decade or so, they too may eventually announce that their students have bolted into the ranks of high achievers.

See how easy it is to improve America's schools?

Vincent Carroll is deputy editorial page editor of The Rocky Mountain News in Denver.

The New York Times

THE WEEK IN REVIEW

Sunday, October 23, 1988

WHERE WE STAND

By Albert Shanker, President
American Federation of Teachers

Getting a Fair Measure of Learning

Testing Needs Regulation

We live in an era of deregulation. It's unpopular to call for regulation and government intervention, but I'm going to do it anyway.

It was not very long ago that people were routinely cheated of their hard-earned income whenever they went to the market. Scales were often fixed to give short weight, and containers didn't always contain what they claimed to. Most of us now shop with confidence because there are government regulations and inspectors. The problem hasn't disappeared—some scales or taxi meters are still rigged—but those who try to give us short measure risk being caught and punished.

We in education also need government regulation of "weights and measures," except that our weights and measures are not pounds of tomatoes or taxi fare meters—they're standardized tests and test scores. What's the problem?

A good example comes from a survey done by Dr. John Cannell of Friends for Education (discussed in this column on Dec. 6, 1987, and April 24, 1988 and by Daniel Koretz in the summer 1988 issue of the *American Educator*). It showed that, according to the reported results on the most widely-used standardized tests, the majority of students in all states were like all the children in Lake Wobegone—"above average."

How could this impossibility happen? The answer is that the "average" on standardized tests is not what most people think it is. If your child, school or district is reported to be above average, you would reasonably think that it means placing in the top half of those now taking the test. Not so. What it really means is that you're in the top half compared to a selected sample group that took the test as much as 5 to 10 years ago! Shouldn't parents and the public have an easy way of knowing that?

Dr. Cannell's survey also showed annual improvements in test scores in states where school people had access to the test items in advance but not in states where there was no such access. Clearly, lots of kids are extensively prepped for the tests, while others take them cold. Shouldn't we know which states, districts or schools have access to test items? Should we regulate such access? Are prepped students who score well "better" than those who are not prepped and score poorly because their teachers may be covering other, richer material? We don't know.

Also, districts may decide to change from one test to another. If the new test is "easier," student scores will likely go up—especially if districts report the "rise" in comparison to the scores on the old, "harder" test (and don't bother to tell you about the change). Or if the new test gets you the same or worse results, districts may claim that the new test was much "harder" than the old one or so different that the kids needed time to "adjust" to it. Shouldn't we have a way of knowing when a new test is used and which test is tougher or easier and in relation to what?

Furthermore, the test scores just don't tell us very much. If you're told that 70 percent of the kids in a school or district are above average, what can you expect them to be able to do? Write a decent letter? Understand an editorial? Figure out the weekly cost of shopping from the supermarket ads in the newspaper? The fact is that you can't tell from the way test scores are reported what students can or cannot do.

One way to remedy this would be to create a National Bureau of Educational Weights and Measures. What would such an agency do? It could:

- approve various tests before they could be sold after checking into their accuracy, validity, etc.
- publish a "critical directory" of tests that would describe and evaluate the major strengths and weaknesses of each available test.
- do the necessary research so that scores on different tests can be "equated" with each other to allow people to know whether the 75th percentile on the Iowa Test is the same as the 48th on the Stanford Test and the 61st on the McGraw-Hill Test, etc.
- find better ways of assessing knowledge and skills than the currently inadequate multiple-choice, paper and pencil tests.
- develop and enforce standards for the use of standardized tests, including test security, updating of norms, reporting of results, etc.
- use government funds to do more rigorously what Dr. Cannell and his group did—survey all states and localities to see how they're testing and reporting the results.
- provide assistance to states and localities that want to simplify, regularize and improve their testing programs and their procedures for reporting to the public and measuring progress, for example, by participating in the National Assessment of Educational Progress (NAEP).
- investigate complaints about tests and testing practices and have the authority to rectify abuses.
- handle complaints from parents, students or teachers about unfair or ambiguous or faulty test items.
- help teachers, students, parents, the press, business and the general public with information that will enable them to understand test results and ask intelligent questions about the tests being used and the results reported.

These are all things that need to be done so that we get honest measures. No one is doing it now. Sure, we generally don't like government regulations and wince at the thought of a possible new bureaucracy. But, in spite of this, I don't see anyone trying to abolish state, local or national agencies that exist to make sure that helpless consumers don't get cheated. The public wants to feel safe from rigged scales or any other device to short-change the customer. It's time to adopt the same philosophy in education.

Chairman HAWKINS. Thank you. I think you've presented your views very well. We have had an opportunity to read your full statement and I think you've done a great job in alerting us to certainly some of the specific things that need to be corrected. So we'll look forward to asking you one or two questions.

The final witness is Mr. Ramsay Selden, Director of the State Education Assessment Center representing the Council of Chief State School Officers.

Mr. SELDEN. Thank you, Mr. Chairman. I appreciate the opportunity to speak to you on this important topic.

The Council of Chief State School Officers is a private, non-profit educational organization representing the 50 states and other extra-state jurisdictions on matters having to do with education.

The Assessment Center, which I direct, is committed or aimed at enhancing the information we have for evaluating the quality and the dimension of education in this country and for executing the states' responsibilities to contribute better information.

We feel that the need for assessment information is critical. We have to have a systematic basis in this country for knowing how we're doing. I don't think it's any secret but the Council has been committed to providing valid reliable state-by-state information on education since 1984. We have been at the front line of efforts to expand the National assessment of educational progress, to provide state-by-state data, and we have been working on other areas in educational statistics to develop a sound useful information base in education so we can gauge our progress.

We are seeing that as the stakes for education increase, it's becoming all the more imperative to base important decision on better educational tests. There are large serious problems with past testing practices. Tests that are widely used in the United States are aimed at low-level skills which send the wrong message to teachers and students.

There is an over-reliance on multiple choice-type items because they are convenient to use and efficient, but which are not effective in measuring important educational skills. In the report, there are testing practices, including the way tests are used in accountability systems that may result in unfortunate consequences for the education system. We need to avoid those. We need to head those off.

But it is our belief that the solution to this problem is better testing practice and that better testing practice is within our grasp. It's not beyond our grasp several years down the road. There are examples right now of several states and local school systems who have developed and are using tests which are much better than previously available.

The States of California, Connecticut, Vermont and the State of New York are administering performance tests in education and serving as a model to other states and school systems around the country. Vermont has committed itself to building educational accountability entirely on a portfolio concept of educational measurement.

The City of Pittsburgh and the City of Portland, Oregon are examples of local school districts who have developed good testing programs where the quality of instruction is not distorted or com-

promised to fit a test. On the contrary, the tests are developed to emphasize and support desirable ends in instruction.

We believe that the efforts that we have completed with National Assessment of Educational Progress in order to make it a suitable instrument for state-by-state comparisons are also transforming NAEP into a test representing good assessment practice. The task of deciding on the content to be assessed in NAEP as it became state-by-state was assigned to the Council of Chief State School Officers in concert with other organizations including local school groups, teachers' organizations, principals, and so on.

This spring and February and March of 1990 in approximately 37 states data were collected in eighth grade mathematics using NAEP. The content measured represented a substantial improvement in the testing of mathematics.

Specifically, just to give one example, the emphasis on problem-solving skills in math instruction and the extent to which math tests address problem-solving skills had been minimal in the past. In setting the content for this assessment, we felt it was important to emphasize higher order skills in mathematics, and, therefore, it was stipulated that even at the fourth grade level approximately 30 percent of the exercises be dedicated to measuring kids' problem-solving abilities. Essentially, new exercises had to be written by the National Assessment of Educational Progress to do that, and they were.

For the 1992 assessment in reading, which will be done on a state-by-state basis at the fourth grade level, we have incorporated a number of features to make the assessment technology more appropriate to guiding instruction. In the first place, the items and exercises that we have specified tap higher order cognitive abilities in reading than typical assessments in the past have tapped.

The students read longer authentic reading passages so that the test is more demanding and more representative of real world and academic reading tasks. Performance from the assessment will be reported in multiple scales or multiple forms reflecting the different purposes and kinds of reading that students do and providing information to educators that's more sophisticated and relevant to their planning purposes than a single scale would be.

In this assessment, over 40 percent of the students response time will be dedicated to open-ended responses. The reason for this is that we set objectives or had goals and ambitions for this assessment that simply were not amenable to multiple choice testing.

The kinds of skills that are being measured here are tapping the students' global understanding of a passage which is best done in their words and tapping students' ability to evaluate and judge and express their opinion about a reading passage such as an editorial which is necessary to capture the extent to which you're training people who can be critical readers for citizenship and intellectual development. That, too, has to be done through an open-ended format because giving an argument to somebody in somebody else's words just won't do it.

We're also trying out new methods in this assessment that are truly innovative and not widespread in other assessment programs and using, therefore, NAEP as a test bed for innovation in educational assessment. This 1992 reading assessment will use a portfolio

method to capture kids' reading activities and classroom reading activities and it will also include an oral reading task allowing us to look at the relationship between oral language development and reading for the first time—an important aspect for reading educators.

It is our belief that we have to have accountability information in education. To head off its ills and misuses, we have to base accountability decisions on better tests. We feel that such tests are within our grasp, but we have to invest in the near future in developing and using these tests to avoid a continuing over-reliance on outdated and flawed testing technology.

I might add that we might also work toward coordinating and knitting together local, state, national and international assessment programs so that the best testing technology is shared among them and so that the least amount of student time results in the most amount of useful information for educational decision-makers at every level.

Thank you.

[The prepared statement of Ramsay Selden follows]

STATEMENT BY RAMSAY W. SELDEN
COUNCIL OF CHIEF STATE SCHOOL OFFICERS
BEFORE THE SUBCOMMITTEE ON ELEMENTARY, SECONDARY,
AND VOCATIONAL EDUCATION
COMMITTEE ON EDUCATION AND LABOR
JUNE 7, 1990

Mr. Chairman and members of the Subcommittee, I am pleased to be here to address this important topic of assessment. I am Ramsay Selden. I direct the State Education Assessment Center at the Council of Chief State School Officers.

The Council of Chief State School Officers is a professional organization representing the commissioners and superintendents of instruction in the states and other jurisdictions. The Council also serves as a forum for states to work together on issues in education of mutual concern. The State Education Assessment Center spearheads Council efforts to improve the information base in education.

Importance of Assessment Information

It is absolutely crucial that our society have sound, useful information on the performance of the education system. It is necessary for us to know how we are doing, and systematic collection of performance data is a major piece of the information needed.

The Council has been at the forefront of efforts to provide better achievement data since 1984, when we adopted a dramatic new policy underscoring the responsibility of states to contribute to a better information base and confirming the value of valid and constructive achievement information. The Council has been heavily responsible for the expansion of the National Assessment of Educational Progress, having identified NAEP as the appropriate mechanism for collecting comparative data on the states, having developed the objectives for the assessment in mathematics and reading as NAEP goes state by state, and supporting the 1988 reauthorization for the expansion of NAEP.

Problems with Assessment Methods

As the stakes for states, local school systems, students and teachers increase as a result of our interest in performance information and accountability, it is all the more important to base educational decisions on sound tests. Past testing practice has been plagued by major problems:

- o overdependence on limited item formats, especially multiple choice;
- o overemphasis on lower-level instructional skills; and

- poor testing practices, including inappropriate preparation for tests, over interpretation of test results, and basing too many important judgements and decisions on too little information or the wrong information from tests.

Better tests are part of the solution to these problems, and better testing practice is within our grasp. It is not just a vague, unattainable ideal. Better tests would:

- use a broader array of more creative exercise formats to do a better job of tapping student performance;

- emphasize deeper subject-matter content and more sophisticated reasoning in those subject areas; and

- be used in ways that avoid inappropriate teaching to tests and inappropriate decisions or judgements based on tests.

That better testing is within our grasp is illustrated by the following examples:

- Large-scale testing programs using integrated tasks, performance items, or portfolio methods have been developed by the Pittsburgh schools, the states of New York, Connecticut, California, and Vermont, the National Assessment of Educational Progress, and the International Association for the Evaluation of Educational Achievement (IEA).

- The National Assessment of Educational Progress in mathematics this year included, as a result of the specifications we developed, a 30% emphasis on problem solving skills, open-ended item formats, and new sections using calculators.

- In 1992, following objectives and specifications we developed for the National Assessment Governing Board, the NAEP reading assessment will:

  - address new, higher-order cognitive abilities in reading;

  - include longer, authentic reading passages;

  - report performance in multiple scales corresponding to different kinds of classroom and real-world reading situations;

  - contain 40% of student response time in open-ended items, asking students to respond to reading questions in their own terms; and

  - try new assessment methods: an oral reading fluency measure, a portfolio approach to reading assessment, measures of students' ability to direct their own reading skills, and an index of reading activities.

The assessment will also provide information on how different student groups are doing, and it will give information on reading instruction methods and resources, so states can begin to determine what may be causing their reading problems or successes.

Responsible Assessment and Accountability

The American educational system must have accountability information. To head off its misuses, we must base conclusions on better tests. But, we must invest in the development of better tests and in better use of tests by educators, to avoid a continuing overreliance on flawed and outdated methods of assessment and misuses of tests by school systems.

We might also work to coordinate and knit together our various assessment programs at the school, local, state, national, and international levels. This way, the best technology can be shared, and with the least amount of intrusion on student time we can gain the most useful information about their performance for each level. We are working this summer for the National Assessment Governing Board on recommendations on this area.

We appreciate the opportunity to comment on these important issues, and I will be happy to respond to your questions or comments.

Chairman HAWKINS. Thank you very much. The Chair would like to open up with several questions.

Mr. Selden, you have indicated the manner in which you think state-by-state testing would be used. If we begin with a test that doesn't really indicate the potential of an individual to perform, whether it's on the job or in education, and then try through state-by-state comparisons to ascertain just where we are, aren't we building in a false system of actually assessing just where education is today? Wouldn't it be rather useless to talk about achieving goals when they're based on such methods of assessing just where students are and whether or not they are going to be first in math and science by the year 2000, or whether or not they are going to be ready for school?

They may, by all of the formal test results indicate that they are ready, and yet not actually be ready and aren't we somehow kidding ourselves?

Mr. SELDEN. No. You're absolutely right. We cannot base state-by-state comparisons on tests that do not tap students' full potential to develop themselves intellectually. The setting of these objectives for these state-by-state testing programs in the National assessment—the one conducted this year in mathematics and the expansion in 1992—consciously set the content of the test slightly ahead of where we know instruction is so we can tap our level of performance in terms of where we want to go not where we are now.

I think the inclusion of relatively a substantial amount of problem-solving exercises in this 1990 math assessment is a good example. That's beyond the attention given those skills in instruction right now, but we believe that's the direction that instruction should go so we have to take the ceiling off. We also have to send a message to students and teachers that these skills are important and the tests are a very important vehicle for sending that message.

I might add, though, that we can't change instruction only by having tests that serve as a carrot out in front of the wagon. We also have to give teachers and school systems support to show them how to expand instruction and move it forward toward more ambitious objectives too. The teachers have to be shown how move toward these objectives.

Chairman HAWKINS. Well, let's say you are administering a test in reading. You may have individuals who are just not good readers, but they may be good performers on the job or may otherwise be good in their classes—but they just happen not to be good readers. So they flunk that test or they are very low. Then you begin to track them into some other lower grade educational experience as a result thereof, and the simple fact is they're just not good readers.

Mr. SELDEN. Well, I think that it's important to——

Chairman HAWKINS. They may have linguistic problems, for example, or cultural problems about reading.

Mr. SELDEN. I think it's important to recognize what Dr. Haney said about using tests to make individual decisions about students. I agree completely with him that no single test should be used to make an important decision about a student that affects their

future or their fate but that much more information should be brought into that decision.

Chairman HAWKINS. But aren't we doing it though? Aren't we nevertheless relying very heavily on tests despite the fact that we don't want to even challenge them and educational progress is being measured that way year after year?

Mr. SELDEN. Well, again, I think there's a distinction between tests that tap an overall level of performance so we can see how the system is doing. We want to know how well kids are learning to read in the elementary and secondary education system and we need tests that can tell us that. But at the same time, Dr. Haney is right. We should not be using any single reading test to track students or to set their fate for a number of years in school process. That's a practice that really has to be curtailed.

Chairman HAWKINS. I guess the important thing is how are we going to curtail it and prepare individuals for the future? We know good and well that 85 percent of those entering the labor market between now and the year 2000 will be minorities, immigrants and others who have grave cultural and linguistic and many other differences than the majority of the students.

Yet they're going to be tested out of current programs because they're failing—they aren't going to get into college. They may get a high school certificate if they get that far—they may be discouraged by the tests prior to that time and drop out. Yet these are the ones that we're concerned about in terms of developing their talents. They may make a good technician in industry but they aren't going to get there because of our testing system discriminating against them.

Mr. SELDEN. I agree with you completely. Let me point out again that the tests that are being administered state by state for the National Assessment of Educational Progress in reading in 1992 will not be reported out on an individual basis. They will not be used to report individual scores.

Most of the tests used in state and local school systems originally were selected to determine the performance of the system as a whole. They were intended to be program evaluation devices. I think what's happened is that in school practice, principals, teachers and other administrators have begun using performance on those tests to make individual student instructional decisions.

Decisions are based on readiness tests administered at the kindergarten or first grade level and then decisions are made as soon as results become available for students on individual achievement tests. I would agree that that is an improper use of test to the extent that kids are classified based on one score. We know those tests are not up to that job.

One other point here is that we have been frustrated that the profession of reading education has really absented itself from the discussion of how we ought to develop assessment techniques in education in reading. I think that the problem that you're talking about really calls for reading education professionals to develop a set of recommendations, recommended practices on the kinds of tests that should be used for student diagnosis and the ways in which they should be used because this is a big problem, a big void

that hasn't been filled, and we really ought to be able to look to that profession to do it.

Chairman HAWKINS. Well, my understanding is that the National Assessment of Educational Progress is coming out in the fall of the year with a new system or with a refined system. They are going to classify achievement into three classes—basic, proficient, and advanced.

Well, I can tell you now who's going to be in the advanced classes. I can tell you now who's going to be in the lowest classification. It's going to be 35 or 40 percent of the students of this country. Yet they are going to be classified into the lowest class and they are going to have difficulties overcoming the stigma attached to that low classification.

Now, I think this is a problem for the Nation. I'm not trying to zero in on the Chief State School Officers. It's a problem for the Nation because here you have an organization, although it may be doing a reasonably good job, it depends for its existence on the Department of Education. If the Department of Education desires a certain particular outcome, that organization is going to continue to be funded.

If it differs with its creator in this instance, it isn't going to get the funding. Yet it relies heavily not on other alternative measures but primarily on a standardized testing system. And next they are going to come out, presumably, with national standards.

Mr. SELDEN. Well, I take issue with the professional dependence of the Council on the Department of Education. We receive funding from dues from our member states. We receive funding from several private foundations. My center is heavily funded by the National Science Foundation. So the Education Department is only one of several sponsors to the Council. We could live and survive and would continue to operate without their support.

In replanning the National Assessment of Education Progress, we absolutely insisted that we would be able to come up with recommendations for the best way to do that without any pressure or constraints from the Department as to the kinds of recommendations that we would come up with.

Chairman HAWKINS. Well, the state departments depend on the Department of Education also for funding.

Mr. SELDEN. Excuse me.

Chairman HAWKINS. I said state departments of education depend on the Federal Department of Education also.

Mr. SELDEN. Well, I think that's true, but that's a whole other matter. I want——

Chairman HAWKINS. I haven't seen one yet that wasn't trying to get as much of the Federal money as they possibly can.

Mr. SELDEN. I think that that's a topic for another hearing. I do want to——

Chairman HAWKINS. May I simply say that I'm not accusing you of any wrongdoing and I'm not in anyway saying that there's something evil or sinister in the operation. It's just human nature. It just seems to me that if everyone profits in a broad sense from a system that protects those who are in the system now and doesn't challenge the system, that you are not going to get what we want in terms of achieving the National goals. The teacher and the class-

room are going to be judged by the extent to which results are obtained according to accepted national standards.

I suppose it gets down to how can we get the independent professional testing system that will be independent of these deficiencies that are built in. Who is going to do it in a professional way and not be dependent on its source of income from some political ideology? I suppose that's the real problem. I don't have the answer. I don't know what we're going to do about it?

Mr. SELDEN. I'll make the recommendations that we made from the 1992 reading assessment available to your staff, and they can review those for their political independence from the Department.

I think we have developed a concept of reading here which is intensely challenging to the U.S. education system, puts it on the line to see how it's performing and is not necessarily in the best interest of states or the Federal Government in terms of trying to make the current system look good.

I think in terms of the results that you anticipate coming out of this 1992 assessment, you're right. There are a lot of kids who don't achieve well in the school system right now and unless a miracle occurs in the next two years, they're not going to be achieving much better by then.

But let me remind you that this assessment is going to be taken on a sample of 2,000 kids in each state in the fourth grade level. That's about ten kids in each of 200 schools or 20 kids in each of 100 schools, depending on how it's done.

None of those kids will be given an individual score. None of those kids will be stigmatized by taking the National assessment in reading. No school will be stigmatized because the schools are part of a state level sample, the kids in the school are not representative of the school itself. So the school won't even be reported out and stigmatized so the teachers in that particular can't be stigmatized. Instead, we're going to get information on the relative performance, the representative performance of each state. That's going to put pressure on the system as a whole to deal with the weak spots in the educational program.

And yes, disadvantaged and minority kids are probably going to do poor on this test, and it's my sincere hope and intention that the comparative testing will stimulate and result in curriculum specialists at the state and local level, in teachers, in legislators, in policymakers looking at their practices and trying to identify effective ways of doing a more effective of teaching reading to kids who do not do well now.

Chairman HAWKINS. Well, I hope you're right. But I predict that in five years, everybody's going to say we're doing wonderful and that we've progressed and that the students are learning and every state is going to maintain that it's doing much better than the average et cetera.

Then we will find out eventually that in comparison, students in other countries aren't doing as well as we think we are doing because we've been mislead by false results in assessing the progress of students. That's what I fear. I hope it doesn't happen, but that's what I see.

I was encouraged by many of the statements in the Report of the National Commission on Testing and Public Policy. I think it's in-

dicated by the fact that I marked up my book so much that there isn't very much left to mark up.

But I think they've pointed out the very dangerous situation that we have now that we built into the system, and that the tests are actually going to have a very negative impact on instruction and that we are still going to rate a certain percentage of kids as failing. Yet, a lot of those kids with a little help—if the test didn't mislead us—with a little help, could be very good. That's very, very true, I think, in the inner cities particularly where we have a lot of immigrants and a lot of kids dropping out of school because they become discouraged.

So I hope that my pessimistic outlook is going to be improved because we have experts like you who are here with us today suggesting some alternatives and being able to use those alternatives as soon as possible.

Mr. SELDEN. Well, let me pick up on something that Mr. Haney said, and that is that we can't used tests that are biased in the effects on students by virtue of their characteristics—their social, their cultural, their economic characteristics.

One of the problems with old style tests is that they reliance on the multiple choice methodology and other kinds of test questions—these are things that the studies on bias in testing indicate that middle class and upper middle class kids are much more comfortable with. Whether they get coached in how to do these things or whether they have more experience with them, that's one of the explanations for group differences on standardized tests.

One of the reasons that we're recommending that 40 percent of this NAEP reading assessment be open-ended is so that we can tap a kid's understanding in their own terms. So that regardless of how the kid expresses himself or herself, when we ask them what does this story mean, we can use their words to make a judgment of what they have learned and what they have understood.

That is, I think, a critical breakthrough in getting around the kind of bias built into unsatisfactory test formats in the past. I think it will be especially important for cultural, economic and linguistic minority kids who have been showing the worst difficulties in traditional test performance.

Chairman HAWKINS. Well, thank you. Mr. Goodling.

Mr. GOODLING. Thank you, Mr. Chairman. I have a lot of the same concerns that the Chairman has. I guess in our drive to insist on excellence, and that's what we are trying to do here in the Congress—instead of just passing pieces of legislation, we're trying to determine how well they're doing and trying to emphasize excellence—my concern is that as we do that we may not have the proper tools to determine whether there is excellence or not as a result of our program.

I always, as an educator, insisted that the teachers use tests primarily for one purpose and that was to determine where the youngsters were doing poorly and then do something about it. I would hope that we would continue to emphasize that part of testing.

My secretary of education in the state from which I come had the great idea that he used tests to rank schools. His purpose for using tests certainly was not fine because he then published a list

of how the schools were ranked throughout the state which, of course, was utterly ridiculous.

He had the great idea that Upper St. Clair was number one in the state. Upper St. Clair should have been one. All the parents are Ph.D.s. They have more money to spend than Carter has liver pills on education. Now, if you take test scores from that area and compare them with those of a school district with a very small taxing base, that is totally unfair.

So my hope is that whatever you design in the future, it will first of all be used strictly to help students improve whatever their weaknesses are and not to rank and rate because I think that's a misuse of tests.

I don't really have any questions for any of you, just a hope that you will continue doing whatever you can to make sure that tests are worthwhile, effective, and measure whatever it is we're trying to measure.

Dr. FALDET. Mr. Chairman?

Mr. SAWYER. Yes.

Dr. FALDET. If I may, I would like to make a comment or two relative to some of the testimony that was given in response to Mr. Hawkins' question. May I be permitted to do that?

Mr. SAWYER. Surely.

Mr. FALDET. One of the statement was made that the fate of a student who scores low on a standardized test—and I'm not speaking of the National assessment type of test, but rather those that are given in most schools at least once a year—the fate of that student is somehow to doom him or her to perpetual educational no man's land.

That is not the appropriate fate for that student. What is the appropriate fate is that based on information from a multiple choice, non-referenced test used diagnostically is to take that student and do something with him or her, or the groups, that is different than has been done in the past.

Indeed—if I may relate just one success story among all the failures we hear about—In 1971, that's a long time ago, a norming was done on one of the major standardized tests and those norms were used for the next seven years. Indeed, in schools using that test, the average of each school began to improve. Now, you could say it's because they familiar with the test or they were teaching to the test.

In 1978, a test was re-normed. We went out and got new students who had never seen that test before. And indeed, in grades kindergarten, one, two and three, performance had increased. Now this was good news, except that when you apply that new standard now to the schools that have been using the test, the percentile dropped.

What brought about this change as we interpreted it? Well, one critical factor was the Chapter One funding that had been going on because this happened in those very areas where efforts had been made to assist those students who indeed where low-scoring students.

So I would argue that means exist today within the schools and within current testing policy and practice to make a difference based on those scores. I think that when the scores, however, become so embroiled in the political arena, it does indeed take

away from the teacher the motivation to do the right thing rather than to teach to the test.

Thank you.

Mr. SAWYER. Mr. Hayes.

Mr. HAYES. Thank you, Mr. Chairman. Let me first apologize for my getting here a little late. As you know, it's common practice around here to do two things at the same time. Therefore, something has to suffer.

I have one question, I guess, directed towards Dr. Haney. You are here representing the National Commission on Testing——

Dr. HANEY. Yes, sir.

Mr. HAYES. [continuing] and Public Policy. After three years of study, the National Commission is now criticizing the use of standardized testing. I believe, too, that standardized testing has been used to weed out people of opportunities. The Commission concludes that under no circumstances should individuals be denied a job or college admission exclusively based on test scores.

Now, my question is, could you elaborate on what other factors can be taken into consideration beside test scores for, let's say, entrance into an institution of higher education?

Dr. HANEY. Yes, sir. A prime example in that case would be a student's previous academic record concerning, for example, both the kind of courses they have taken and the grades that they have received in them.

There are certainly problems with grading practices in our secondary schools across the Nation. But quite consistently over a period of 50 years research has shown that students' high school record actually predicts their subsequent performance in college better than standardized college admissions tests. Moreover, evidence clearly shows that if you did rely more heavily on students' academic record in high school for college admissions than on standardized college admissions tests, that would result in less adverse impact on groups of individuals who tend to be particularly adversely affected by decisions based solely on standardized test results, such as, individuals from African-American backgrounds, individuals from Hispanic backgrounds, individuals from poor socioeconomic status homes.

That's a very practical example with regard to the question you raised, sir.

Mr. HAYES. Can you maybe speak just briefly to the needs of the year 2000 labor market? How the current use of standardized testing may negatively impact on our readiness for competition which we have so often alluded to?

Dr. HANEY. Yes, sir. I can give you some concrete examples that were brought to the attention of the Commission and then try to speak to you briefly about what the Commission recommended to try and to remedy those kinds of problems.

One example. When the U.S. Employment Service was trying to development a new referral system that wouldn't be as expensive, they started experimenting with a referral system for people who go to the Employment Service looking for jobs which would be based exclusively on a test called the General Aptitude Test Battery. It had been used for employment referrals based on some theories that I won't try to recap here.

But what they found was that in some communities when they started placing exclusive evidence on this standardize test for the purpose of employment referrals, some groups of individuals—disadvantaged individuals, unemployed—simply stopped using the Employment Service because they felt they couldn't get a fair shake on this test.

So the use of the test was clearly undermining a prime objective of the Employment Service to try help place people who are unemployed in job. That's just one example of how over-reliance on standardized employment tests can undermine vital employment policies.

That kind of issue is going to be increasingly important because, as I believe Chairman Hawkins alluded to, we're going to be having vastly increasing proportions of the entry-level work force in the next ten years composed of minorities and women than has been true in the last 90 years.

Now, that suggests to the Commission that what we've got to do is to try to avoid, not just an education, but in employment selection practices, over-reliance on just one form of evidence. There's been considerable research, particularly on employment assessment, that relying on alternative kinds of assessments than standardized tests reveals not only equal validity with standardized test results but also smaller adverse impact on the sorts of groups that have historically been disadvantaged in our employment system in the past.

Our full report does point you to some of the examples of those kinds of alternative assessments and the evidence concerning their validity and lesser adverse impact.

Mr. HAYES. Thank you. Does anyone of the other members of the panel want to comment on that question looking ahead to the year 2000—our readiness, so far as the labor market is concerned? Are you satisfied with the response that I received from Dr. Haney?

Mr. SELDEN. Well, I would add that apart from the employment identification or screening practices which Dr. Haney's addressed, having a competitive work force in the year 2000 also depends on having an effective school system. Having an effective school system in my mind hinges in part on having valid, useful information on how kids are learning.

That's going to require better tests and better use of tests between now and 2000 and monitoring the system to make sure we're getting enough students who can do what our labor force needs and to make sure that all students have an equal opportunity to prosper in the education system and to join that labor market.

Mr. HAYES. I know I've exhausted my time, Mr. Chairman, so go ahead. Thank you very much.

Mr. SAWYER. Thank you. Mr. Petri.

Mr. PETRI. Thank you. Thank you. gentlemen, for coming here today.

When you talk about tests, are you talking about testing for knowledge or ability? I mean, what is it that these tests are designed for? There's a difference.

Dr. FALDET. May I respond to that.

Mr. PETRI. Yes, because there's a difference between testing what people know and their ability, and it would seem to me that one of

the objects would be to see if there is a difference between a person's level of knowledge and his level of ability. This is a separate question from the fact that when people take any test there's a continuum and some people do well and others do badly. That's in the nature of things. If everyone does 100, forget it. But if there's a big gap between ability and performance, for example, then the test can maybe show that, and you have a person who has a potential that has not been realized and we ought to do something about it.

If the person's ability and performance are both lousy, well, okay. Or if they're both great, okay. But when there's a difference between the two levels, then the test has revealed something that shows that we're not helping people. And that seems to me the value of tests, really—not to disguise differences between people or point them out or anything else.

Dr. FALDET. I would like to respond to that just a bit because I would like to remind the committee that we're talking about several different kinds and levels of tests here today. There are those tests for college admissions, the SAT, the ACT, the National Assessment sampling type of testing in terms of the people that are contributing to the data. And we're talking about employment testing.

But I think one of the key problems that all of us recognize is that testing which goes on in elementary—particularly elementary—and secondary classrooms. It's that testing that is not a pass/fail kind of situation. It is indeed on a continuum. But within that time spent in testing, which may be a day and a half of time, you get potentially a tremendous amount of information. And that can be both with respect to some measurable abilities as well as some measurable current levels of skill.

It's that kind of information which needs to be acted on. You want to make sure that you use test information in instruction. I think we get very fuzzy when we are afraid to use test information about an individual to say I need to do something differently with you, or with you two or three students because you do not yet have the basic skills that will permit you to do reading problem-solving at some later point.

We very recently have had a great deal of concern about back to the basics, and have we forgotten that? Are we no longer concerned about those things which are commonly measured in standardized tests? I would invite the committee to look at any elementary test battery and ask yourself what is it that is being measured here that I really am not concerned about students knowing or having been acquainted with.

If indeed we are abominably failing in teaching these things, as shown by standardized tests, or even if we are teaching to them, I think the problem needs to be addressed in addition to those problems where misuse of test results for political reasons, pass/fail, employment decisions, are also an issue. But it is a different kind of issue, and I would not like to see them mixed.

There's a lot of talk about how we are going to be prepared as a country economically. But it seems to me as a government, we have another big concern and that is that we have a literate community of people prepared to exercise the responsibilities of citizen-

ship in a free society. This was always one of the great rationales for having public support of universal education, so that everyone would have an opportunity at least to learn how to read and write, read the newspaper and follow events, and contribute to a democratic society. That's the basic rationale for public education in my mind anyway.

Businesses and other people will be able to figure out ways at the end of the day to impart skills so people can be productive, but they don't necessarily have an interest in preparing individuals for the responsibilities of civic participation.

In that connection, we each think of our own background, I suppose. I remember a teacher I had who would make us get a score of 90 or above when we were seniors in high school on the parts of the English language such as adverbs, nouns, verbs and so on. If you got below 90, as far as she was concerned, you came after school for an hour. The purpose of the test was to raise everyone in the class to a certain basic level. She didn't feel you should graduate from high school if you did not understand the basics of the English language and how to put a sentence together and all that sort of thing.

So people were just given extra instruction and they kept on coming in until they all got above 90, even if it took a week or a month or two months on a particular part of the exam—and there must have been about 50 different exams that we had to take in the course of the year.

So it seems to me that rather than one snapshot of someone's performance, the test can be used as a guide for helping people to reach at least a basic level of competence which we want to encourage and expect all people to have if they're to be participating citizens. Is that at all a valid approach, or is that a waste of our time?

Dr. FALDET. No. I think that is what testing in the schools is all about. It is a first step in early identification hopefully of students that are having some difficulties in some areas that are agreed by the school and generally, I think, nationally are important things. Things that are basic to subsequent learning.

They are not the end of what should happen because there should be confirmation. I think there are students that have bad days on a standardized test or a custodian is mowing the law right outside the window. Yes, that can happen, obviously. But you first of all confirm and say, yes, this is consistent with what behaviors I've observed and now I have some evidence to enables me to confirm my thoughts and we're going to do something about this.

That's were you develop then, as a part of a total assessment evaluation plan, what your next steps are. I think that is key to any, if you will, guidelines that you would put out in terms of use of test information. It's got to be in the context of it being a continuing thing and that actions are taken as a result of it.

Mr. SAWYER. Let me interject, if you don't mind. We've been talking about a couple of different fundamental differences in the use of testing. One of these is the notion of the gatekeeper, the portal through which everyone must pass—a right of passage approach to testing. The one that is far more complex and useful in the longer run is testing early and often for diagnosis and then targeted remediation.

The frustration that I have is that as we talk about gateway procedures, we speak to the inadequacy of current testing procedures to be used and the consequence of stigmatizing whole population segments, schools and school systems, and individuals, is a matter of deep concern for all of us. If these instruments are inadequate for that broader less precise purpose how do we move them as tools into the realm of early diagnosis and targeted remediation which I'm sure we can agree is a preferable approach?

Dr. HANEY. Could I suggest a very quick answer to that by coming back to Mr. Petri and asking a question about your teacher—your high school teacher who did this testing with regard to grammar. Did she—after you took the test what happened? Did she give it back to you?

Mr. PETRI. The ones that we got wrong we were told about, and then we had to take another test the next day.

Dr. HANEY. Right now because——

Mr. PETRI. She'd give it back to us. Sure.

Dr. HANEY. She gave it back to you.

Mr. PETRI. Yes.

Dr. HANEY. Right now because of the nature of most testing technology, including the commercially published standardized tests that Dr. Faldet was talking about and the National Assessment instruments, which I think on many counts are quite good, that does not happen because you cannot give students immediate and detailed feedback on what they learned without invalidating those items for future reuse.

Thus, when the schools have been under a lot of pressure to improve test scores, you find exactly the kind of Lake Woebegone effect that our third witness talked about today. I am sorry, it's Doctor—or Mister?

Mr. FAITHORN. Faithorn.

Dr. HANEY. Mr. Faithorn spoke about with regard to his colleague's Dr. Cannell's useful work.

So that I think we have to be careful, number one, to distinguish between different kinds of testing for the purposes of allocating opportunities at major transition points in peoples education and employment careers. They are really instructionally aimed. For instructional purposes, you want them related to what's being taught, you want to be able to provide feedback and so you are talking about fundamentally different kind of testing right now.

Mr. SAWYER. Well, not necessarily. We're talking about test instruments that test kids in fourth, eighth, twelfth grade. To get to the twelfth grade, or maybe even eighth, they become rights of passage. But even if you wait until the fourth grade, it's too late to do the kinds of things you're talking about.

Dr. HANEY. It's too late.

Mr. SAWYER. And that real diagnostic instruments need to come early and often and became the kind of tools that will help teachers target their efforts and approaches. Perhaps that strikes people as a little brick schoolhouse approach, but it worked for a long time.

Dr. HANEY. Well, I think that some new testing technology is going to make that more possible so that you can give people detailed results without invalidating the test for future use.

Mr. SAWYER. Let me move to my second question then because in doing this we've talked a great deal about changing the structure of testing. As we move to particularly more open-ended answers, and answers that require subjective analysis of a response, how do we go about standardizing the quality of evaluation? Has there been much thought given to that sort thing, gentlemen?

Mr. SELDEN. Can I respond? I think that we are demonstrating in tests that are used on a wide scale basis that open-ended responses can be scored validly, accurately and reliably.

It think the best example is in writing, we used to have multiple choice writing tests in education and now a number of states and a lot of local school districts and I think there are several commercial——

Mr. FALDET. And publishers——

Mr. SAWYER. Well, my friends who take the bar grumble about that all the time. I don't know, I've never done it.

Mr. SELDEN. It's a matter of setting criteria. Given how a person responds, you have to preset criteria for what you are going to deem an acceptable or unacceptable response and then people can be trained to score the responses and judge whether they're correct or acceptable or high in quality and low in quality. And that's done.

The state of New York has a fourth grade science test where kids come up to a table and they actually conduct an experiment in order to find out what shape an electronic circuit is. They are watched while they are doing this and the teacher judges whether or not they successfully designed an experiment and carried it out to do it. But it's an integrated task, it takes a certain amount of time for the kid to do. Many kids may do that in different ways. But it was administered to every fourth grader in the state of New York.

Dr. FALDET. Mr. Chairman, indeed, you are right. There are certainly available now some writing sample kinds of test together with keys for scoring. I think that you also find some portfolio concepts available commercially.

I think the important thing is to recognize that whatever change there is, you are going to have to convince and involve those local educators without whose commitment, understanding and support, whatever is legislated or developed is not going to get implemented appropriately, and I would suggest we're seeing that now.

I think we were in a period when standardized where used far more appropriately than they are today. I don't think they were any less representative of the curriculum that we wanted. But there have been pressures that have created now high stake situations revolving around that and I would urge you to make use of the expertise of those who deal daily, weekly and annually with schools in assistance in designing how this is going to be done and how you introduce it and implement it.

Mr. SAWYER. Thank you. Mr. Smith.

Mr. SMITH. Thank you, Mr. Chairman.

Mr. FAITHORN. Mr. Chairman?

Mr. SMITH. Let me, if you could because I am going to have to go in about 10 or 15 minutes and maybe your question will fit with my request, if that's all right sir.

We are on the verge, I think, of making the same mistake we make every time we talk about how to make schools better. It is an a point, with all due respect, of convincing or just involving teachers and school boards to use existing or new technology. It is a question of asking them what it is they as professionals would do. If we want to make diagnosis the basis of how we determine how much value we are adding to children's lives cognitively and in terms in skills and behaviors and attitudes, then, in fact, it must start with the school, not end with the school.

Paulo Friere would not be welcome, in my mind, in any of our schools. Yet he still has—especially for children who do not share the so-called dominant culture of our country—he still has for my money the single best philosophical instructional approach to teaching reading, which is to take the words of power in the culture from which you come and use them to pull the child to learn how to interact verbally and in reading and writing with a culture that he or she is going into.

Somehow, if we want education, which is derived from the words ex ducare—to lead from—to lead beyond, if that's really what we're serious about, we need to figure out a way to blend the social imperative of schools, which is a common socializing experience which binds our culture together on it's good days, and then the notion of excellence, which is that we maximize the capacity of every student.

From my point of view, that means that every time we use a test simply to judge, it is an external operation determined by somebody else and it in fact by definition has to be destructive to the educational process which would be based on diagnosis and evaluation which would involved not every three years but hopefully twice a year or more cogent comments about how well a student is doing and what that individual knows or can do differently than he or she could do six months before so that parents can understand it and the community can understand it in relationship to what their goals are for the student.

I think it speaks strongly for the idea of flexibility in our schools so that we could teachers—I happen to have a bill which does that——

[Laughter.]

Mr. SMITH. [continuing] so that we can ask teachers and boards how it is they would like to organize an educational program so the capacity of every child is maximized and can be described in real terms.

Two questions. One, and these may—I think there's an economic plot here too. I wondered to the extent any of you have investigated or have opinions about the connection between the textbook industry in this country and the testers because the last time I looked there's big money on the table and they go down in Texas and California—with all due respect to some members of this committee—involves a whole lot with how it goes down in the other 48 states because there's a lot of kids and books. I think there's an unholy relationship, and if we don't understand the economic impact, I think consequences of reforming testing, we're never going to get at it.

Secondly, the question of whether a little diversity and how we evaluate and describe learning—not saying there's any one way, but letting states and schools go at it differently for a while until we find out what the good practice is and let it bubble up.

How do you feel about diversity and how do you feel about the testing and text alliance in this country?

Mr. SAWYER. You each have 30 seconds.

[Laughter.]

Dr. HANEY. Very briefly——

Mr. SAWYER. I was kidding about that.

Dr. HANEY. Okay. I could very briefly, in 30 seconds indeed, say that I think your concern about the changing nature of the test and textbook publishing industry is right on target because there has been a tremendous number of acquisitions in both the textbook publishing and the test publishing industry over the last 10 years. I can't cite them off the top of my head, but I can provide you with some documentation of that.

Mr. SMITH. Please do.

Dr. HANEY. Not only, though, must we worry about test publication and textbook publications, but also there now is questionable practice in people who publish tests, publishing test preparation manuals for those tests which appear to have been adopted fairly widely in some schools at substantial cost. So I'd say that is a concern that is salient right now in light of mergers that have happened over the last five to ten years.

With regard to diversity, I think you're absolutely right. There is considerable evidence on the basis of assessments that have been made in the past and studied through research that when you start using different methods of assessment, you start beginning to see talent in different people and in different groups in different ways.

There is a tension, as you alluded to, between trying to educate people and trying to make judgments about them. To the extent that we want to form educational decisions based about people and students—particularly young students—in context, we have to rely, I think, more on the nonstandardized evaluation systems that grow out of the local context because there has been research that shows that things as seemingly innocuous as to whether or not students have had breakfast in the morning, can significantly affect their standardize test results.

There's not way that the companies that Dr. Faldet represents would have any way of knowing that when they score the test results. You'd have to rely on the teachers who know the students in context.

Mr. Faithorn. Mr. Chairman?

Mr. SAWYER. Mr. Faithorn.

Mr. FAITHORN. I'd like to respond to Mr. Smith's question and touch on Mr. Petri's comments also.

Mr. Petri was talking a teacher on one end of a board and a student on the other—the ideal learning situation, if it's a good teacher. The testing that we're upset about and Friends for Education doesn't provide any feedback—and with respect to Mr. Smith's question—it involves really serious money. What Mike Royko in Chicago would call really serious money—the cost of these standardized tests and the textbook that goes with them and the prepa-

ration books that go with them and the relationship between the publishers of these tests and the school boards. It's big business.

With respect to Mr. Hayes' question earlier—Mr. Hayes earlier asked a question about going into the year 2000, what ideas did we have with respect to that. I would like to respond to that kind of indirectly by saying that I went around to the Department of Education and met with officials in their Undersecretary for Research Office to better prepare myself for this first time I've every been before a congressional committee.

They confirmed the fact that they had checked out our study about the phoniness of standardized test results and felt we were right. But they said that they had not checked on any of the implications about cheating, our allegations about wholesale cheating that's going on in the schools to make the student look better and the school look better in passing these standardized tests.

They said the reason they hadn't done that because this was anecdotal and therefore it didn't lend itself to any real verification, and furthermore the Congress and the school districts didn't want the Department of Education messing around in their affairs to the point where the Department was examining into procedures in schools, where comparing state to state, or school board to school board. That this was a nightmare to all these people and that the Department of Education should damn well stay out of it.

I come to my point in answer to your question. I was appalled by this and I think the Department of Education ought to damn well be getting into questions like that if we are going to do something between now and the year 2000 in closing the gap between our kids and the other Western democracies and industrial states of the world. Thank you.

Mr. SAWYER. Mr. Poshard.

Mr. POSHARD. Thank you, Mr. Chairman. I'm sorry I got here late, I had some other committee meetings this morning also and so I didn't get to hear the original testimonies. My question to Mr. Faithorn is why would we even need to cheat if the test are developed as you have explained in your testimony, and I'm assuming you have some evidence through the study in which you engage and so on, to show that they are.

It seems to me that if the norm group is just a group that is tested cold and then that's compared against students who have studied material to take this test for a whole year and the differences are compared, why would you need to cheat. Why wouldn't we come out above average on everything?

Mr. FAITHORN. Well, you're quite right and Dr., this gentlemen on my right, explained just why the performance improves every year when it's compared against an old norm. There isn't really need for cheating but it goes on wholesale anyway. I don't know if you saw CBS's 60 Minutes——

Mr. POSHARD. Yeah, I understand that, but that's not the question. I understand that there is some cheating, but I'm more interested in the other facet here. Mr. Faldet would you elaborate just a little bit about Mr. Faithorn's statement of the way the norm group is established? I'm sorry if this has been asked already.

Dr. FALDET. No. I don't think it has. Certainly. The goal of setting a standard for a period of time is too make sure that it's repre-

sentative geographically by ethnic groups, socioeconomic groups and so forth, large districts, medium/small districts—so some rather elaborate strategies and techniques are used to seek the cooperation of randomly selected districts throughout the United States in taking a test for which they will not receive any scores because there really isn't anything to report back to them at that time, and that can influence the level of motivation on the test.

But from those studies a variety of things come. Certainly, assigning the percentiles—what represents the 99th percentile, what represents the median—the 50th percentile by grade and semester. In addition, that's where you get the reliability and the beginning of validity studies that have to accompany each standardize test, but then you begin to give it to people who have not seen it before, but who have chosen this test hopefully because the objectives measured are as consistent as possible with the objectives that their district is emphasizing. I think that's key. Then they begin taking it and then indeed the scores may begin to rise

Now I don't know how much of that is because teachers are teaching to the test or teachers are indeed to continuing to emphasize the objectives that the test is measuring. In the latter case, I think it would be good. In the former, it's abominable.

Mr. POSHARD. I'm assuming there's both pre-study test and post-study test. Right? You're not talking about giving this test one time to a group of students and establishing a norm group. Right? You're talking about giving the test before school, having a full year of school for the norm group and then testing after the year of school. Isn't that the way you establish the norm group?

Dr. FALDET. Yes, sir. It's given generally in the fall and it's given again, probably alternative form to get some variety there, in the spring so that you have some pretty good data on what growth has gone on.

Mr. POSHARD. Okay. Then my question is to Mr. Faithorn. In your testimony you described this group of students upon which the norm is established as taking the test cold. What do you mean by that if the students take it before study and after study, how are they taking it cold? I'm trying to collaborate data here so I understand this. Why would you say they're taking it cold if in fact they have spent a whole year studying the material?

Mr. FAITHORN. Well, first of all, let me apologize for the kind of casual and sloppy language that I used. I thought that this would convey the idea of what I understand goes on which is that a new test is developed by McGraw Hill, let's say. They give it to a group of students and they get the results and then those results are used for the next several years against which to measure subsequent groups of students taking the same test.

Mr. POSHARD. But you couldn't establish a norm if you didn't have subsequent study after the pre-test and then a follow-up test to see how much the student learned. Otherwise you don't have any group to test it against. I mean the two groups that are tested have to have the same experience or else there's no validity or reliability to the test.

Mr. FAITHORN. Well, may I defer to my new friend here on my right to answer that question because it's his business.

Mr. POSHARD. Well, no, but I'm trying to find out what is actually happening. You're saying they're taking it cold. When I read your statement, I thought they're giving this test one time to student and that's it and then they're going out and letting other students study a whole year for the material and take the test. There's no reliability or validity to that sort of procedure if that's what's occurring.

Dr. HANEY. Could I interject——

Mr. POSHARD. My question is does your norm group that you're establishing take a pre-test, study like every other student that's going to receive the eventual grades on this for a full year and then take a post-test and you compare the results for the local district against that norm-referenced group? That's all I want to know.

Dr. HANEY. Yes. I think I can help illuminate this in that I've talked with Dr. Cannell, Mr. Faithorn's quick friend about this several times. The distinction is that when most publishers norm tests, they seek to develop empirical norms both in the Fall and in the Spring.

They choose school systems so as to try, as Dr. Faldet explained, to try to have a nationally representative sample of school systems all across the Nation.

Mr. POSHARD. I understand that.

Dr. HANEY. And they develop the norms from those testings from both the Spring and the Fall. However, when they go to sell those tests, school district studies have shown typically select between the big test series on many grounds, but primarily on the basis of whether or not the test seems to match the local curriculum.

So when the results are subsequently reported you are in effect getting results based on a self-selected group of school systems who may have picked that test because there's a better match between that test and their curriculum. But the norm group was not selected because of any such overlap between test and curriculum so in that sense you are talking about two quite different groups.

Mr. POSHARD. Yes. Okay. Then I understand that fallacy——

Dr. HANEY. One other sort of research finding that may interest you and that I think that your question was an excellent one because while there's been a great deal of publicity to issues of cheating as a result of some of Dr. Cannell's and the 60 Minutes program, a very interesting research report that came out just a few months ago a national survey of teachers and school administrators asking them about test practices.

The results were treated anonymously so the respondents had no reason to cheat, but the results indicated that these people—both teachers and administrators—perceived there to be on the order of 10 percent or less of their colleagues who might have engaged in improper test preparation or what might be called cheating.

But in fact the results indicated that more than 70 percent from the systems from which people responded had engaged in what has come to be called test curriculum alignment so that they had aligned their curricula to better address either the objectives covered by the test or the actual items represented on the test.

The problems is that they were not normed originally on schools whose curricula were so aligned and there is some research on the

ramifications of test instructional overlap on the results. Basically, to try to summarize a fair amount of literature very quickly, it shows that differences in test instructional or curriculum overlap could easily account for the magnitude of Lake Woebegone effect that Dr. Cannell found.

Mr. POSHARD. I'm sorry, Mr. Chairman. One quick question—so then we can be assured that the publishing companies are in effect carrying out correct procedures in terms of norm-referencing their test in regard to validity and reliability. In other words, we're not measuring against a pre-test, and a post-test. Right?

Dr. FALDET. No, sir.

Mr. POSHARD. Okay. That's good and I accept your explanation of the schools actually trying to align themselves in terms of the particular test that they give to the students. Thank you.

Dr. FALDET. But if I may, Mr. Chairman, if they didn't do that, I would be disappointed. If they found after the first administration of a new standardized test, that their students were woefully weak in some language arts skills, and they didn't align their curriculum to correct that situation and thus increase the scores hopefully the next year, I would be disappointed. They wouldn't be the doing the instructional job that the tests are helping them to do.

Mr. SAWYER. Mr. Payne.

Mr. PAYNE. Thank you, Mr. Chairman. I will pass since I came in late and allow my colleagues to—if there are any other questions. I just might make a statement that my opinion of standardized tests in general that we do find that in urban areas this new way of testing has been introduced only more recently in urban areas than what we're able to ascertain that for many years standardized tests were taught as far as pre-K right on up how to take the tests and therefore the natural results are that those who have been trained to take those types of tests invariably would do much better by virtue of their preparation to do that.

I question where education begins and proficient test taking leaves off and there is, it seems to me, you know, in the environment of teaching, where you develop concepts and so forth by just practicing standardized testing. It just appears to me that there's an absence of education.

Of course, there has to be a way to test what has been taught, but I've been somewhat concerned through the years of testing since much of it, as we all know, tends to be culturally biased. I just wonder how you might truly be able to test a really intelligence quotient of a person who has not been exposed to the bias that these test take by virtue of the manner in which they are written or prepared.

So I certainly do not put too much stock in the testing of intelligence, ability to learn in the results of standardized tests. I've seen these types of tests exclude minorities through the years whether it was for employment—at one point in a very large company, I hired a person who through a summer program as a teenager who took the employment test and failed the test for normal entrance into employment with this extremely large form therefore the individual would have been unable to work in that company.

But we, through a back door method, I guess, I was able to continue this person on from a summer program and it was not only

that this person became proficient—now this is a person who would have been excluded from a very simply and basic test at that time.

This person not only did well but went on to become the supervisor, went on to open a department at a new regional home office eight or nine years later. The interesting thing that this individual who is still currently maybe in her middle thirties, early forties perhaps, is still moving up the ladder.

That company to this day doesn't know that she's the one that failed the test. I might even at one point see if I can find her again and maybe discuss some of these situations with her as it relates to the fact that she would have been unable to work for that corporation based on that test.

Therefore, that test had no relevance or ability to perform and achieve. So I, as I indicated, I missed the testimony, therefore I will not ask any specific questions. I just thought I might share those feelings with them.

Mr. SAWYER. Mr. Chairman, you had a question.

Chairman HAWKINS. May I ask Mr. Faldet a question because I was reading his prepared statement. On the bottom of page 6 and the top of page 7, in effect I was trying to see how the actual test is constructed. It is my understanding that what happens on the standardized norm-referenced test is that it's designed in such a way that the National bell curve will result in 50 percent in effect of those taking the test will pass and 50 percent will fail.

Dr. FALDET. No. No, Mr. Chairman.

Chairman HAWKINS. Would you then correct my understanding.

Dr. FALDET. Let me correct that impression if it is there.

Chairman HAWKINS. Well it's also in the National Association of Secondary School Principals book that I see on testing. They say that also.

Dr. FALDET. The only thing I would want to correct is the pass/fail. All of us, no matter on what traits we might be measures, someone is going to be the one that scores the highest and someone is going to score the lowest. That does not imply that even that person scoring the lowest has failed. It just describes. That is the measurement concept. The concept of passing or failing or good or bad comes only after someone puts a value on a particular score.

For example, all of us would like to see every student in the United States scoring above the 50th percentile on every test. Unfortunately, by definition, that will never happen. As the track gets faster, the percentiles change and we say alright there is a new average. There is a new median. It's not a pass/fail. That's an evaluation.

Chairman HAWKINS. Well, let's not use that. Let's say 50 percent will score above and 50 percent will score below. Well first of all, you conduct field tests as I understand it. Then you use the test scores and you eliminate those that everyone got correct and everyone got incorrect. You select out of tnat number of questions those that are not all together one extreme or the other and then according to the bell curve, 50 percent then are expected or graded as above that norm and 50 percent of them are below the norm.

Now there's no assurance that any further interpretation is going to be put on that test. That is you indicated and indicated correctly that a lot depends on how the test is interpreted and cer-

tainly that's true. But is it not true also that, in effect, we already know approximately who's going to be above and who's going to be below that average. We can pretty well predict that below the average, there will be those students who because of language difficulties or cultural differences or varying adverse economic conditions are going to be in that below the average number.

We also know that the children from the more affluent families with parents where they learn answers although they haven't taken the test but they learn the answers from their parents, from their home environment. We know that. We know pretty well that's how the standardized test is going to come out.

You say there should be in-service training for the purpose of correct interpretations and that's right. But we also know that in-service training doesn't take place ordinarily, that those kids who are termed, in effect, low achievers are going to be stigmatized obviously unless it's accompanied by some other measurement, they are going to be classified and rated and forever be subject to that low achievement expectation.

That's a normal situation. It's not your fault. I'm not accusing you of anything, but isn't that in reality precisely what takes place.

Dr. FALDET. That is potentially the fate of someone who scores at the tenth—fifteenth percentile—the lower scoring student. If you were going to predict where that student will be the next year on the appropriate test the following year, you would predict that that's where they would be then if there is no intervention, and all I'm suggesting and what our interpretative materials suggest is that if you have this information and don't do something about it, then you might as well not have the information because indeed you could predict that score as well from the area the student comes from, the socioeconomic class and so forth.

The information is provided not to confirm that indeed low scoring students probably come from more deprived neighborhoods, et cetera, but to identify and confirm those areas in which that students needs some special instructional assistance to, in effect, beat the prediction. That's why we do screening tests in medicine.

It's not to confirm that yes, you have high blood pressure. So long Charlie, you're dead. But to take actions appropriate to remediate, to confirm certainly further diagnostic tests, but to make a difference. That's where I think efforts that might be suggested through guidelines say look we want to know what test you're going to give but we also want your strategy and your ideas and your commitment to do something about it when the scores come back —whether it's a local test or a nationally prepared one that indeed has some other potential values.

Chairman HAWKINS. I'm not accusing test developers and I'm not accusing the state. It may sound that way, but isn't the current education policy driven by test scores and not by intervention and not in-service training and not by teacher development. For an example, when the Secretary of Education calls the schools terrible, as he recently did, in effect he's ignoring what can be done to correct the very situation that the test scores seem to generate not because they're wrong but because we don't follow up.

I think that's what we're trying to do is see how we can best use test scores in the proper way and not as we do now. But we never get around to finding the money for intervention for example.

And so well, I think we agree on at least the implications even though we are very slow in getting the solutions. Thank you.

Mr. SAWYER. Mr. Chairman, you told me when you asked me to take the Chair that I had to get you guys out of here by noon.

Chairman HAWKINS. Oh, I'm sorry.

Mr. SAWYER. I just want to take the prerogative that you've given me and the Chair to say thank you for an extraordinary hearing—one of remarkable importance and one who's topic I hope we can visit again.

Chairman HAWKINS. Thank you for a very remarkable group of witnesses. Thank you.

Mr. SAWYERS. If there is no more business to come before us, we stand adjourned.

[Whereupon, the subcommittee was adjourned.]

[Additional material submitted for the record follows.]

Friends for Education, Inc.

600 Girard Boulevard N.E.

Albuquerque, New Mexico 87106

(505) 260-1745

Working For Accountability
In Public Education

John Jacob Cannell
President

The Honorable Augustus Hawkins
Chairman
Committee on Education and Labor
U.S. House of Representatives
B-346C Rayburn House Office Building
Washington, DC 20515

June 16, 1990

Dear Chairman Hawkins:

Thank you for asking me to testify before the Subcommittee on Elementary, Secondary, and Vocational Education. I regret that I was unable to come to Washington to personally testify, but I believe our Washington representative, Mr. Walter Faithorn, ably presented our organization's views to the Subcommittee. As per your request, I hereby submit the following written testimony.

My views on testing in American public schools are expressed in detail in The "Lake Woebegone" Report: How Public Educators Cheat on Standardized Achievement Tests, a copy of which I enclose. In addition, I enclose copies of a few of the many newspaper articles about the "Lake Woebegone" cheating scandal, as well as a videotape of recent NBC and CBS coverage of the scandal.

Personal Experiences

My views of current testing practices in American schools are colored by three personal experiences. The first is my experience treating adolescents patients over the years, mostly for self-esteem problems. As a general physician, I saw child after child, from upstanding and caring families, damaged by our school systems. Time and time again, I saw functionally illiterate children moved through the public schools like so many cattle. These school's lack of standards stood in sharp contrast to the high "standardized" achievement test scores the school administrators routinely released to parents and the press.

As I became increasingly suspicious of the public school's testing programs, I started sending many of my patients for outside achievement testing by independent testing experts. I found many of these children tested well below grade level on independent achievement testing but both they and their parents were being told they were achieving "above the national average" by school officials. Needless to say, most of these children came from disadvantaged backgrounds.

The second event which colored my view of testing occurred when I queried the U.S. Department of Education about the commonly used standardized achievement tests. I mistakenly assumed, like many Americans, that some branch of the federal government attempted to verify the accuracy of the commercial achievement tests our children take in public school. After all, they are the product of commercial, for-profit corporations that sell and transport goods and services across state lines. I was shocked to learn that the U.S. Department of Education makes no effort to verify the accuracy of these tests. Unlike testing in any other country in the world, the achievement tests given to American children and reported to American parents are not regulated, verified, or overseen by any agency, private or public. Instead the policy of the U.S. Department of Education seems to be: "Let the children beware."

A final incident convinced me that a substantial number of American public schools are releasing falsified achievement data to parents, taxpayers, and the press. After becoming increasingly troubled, I decided to telephone a major test publisher and present myself as a superintendent of schools from a small southern Virginia school district who was interested in buying one of their tests. I called and explained that our board of education was considering changing tests, and the members were very interested in improving the district's test scores.

Almost immediately, I was talking to a saleswoman who implied that our district's scores would be "above average" if we bought one of their tests! She further intimated that our scores would go up every year, at least until we changed test questions.

How could she know that our district would be above the national average? The district whose name I used is a poor rural southern Virginia district. How could she be sure our scores would go up every year? She couldn't know if our district's schools were improving or not.

I had been aware of rumors about cheating in the schools. Many teachers privately told me that school personnel changed students' answer sheets after the test, gave students more than the allotted time, used the exact test questions to review for the test, or made copies of the test to give to their students. Many teachers complained that administrators forced them to teach items known to be on the test, claiming they could not get a promotion without producing high test scores.

It became clear why the saleswoman could guarantee scores would go up every year as long as we didn't change test questions. The schools and the publishers they had under contract were jointly claiming that scores were improving because schools were improving. The schools, in cooperation with their contract publishers, were teaching the students

the answers before the test was administered, and then the districts reused the same test questions year after year.

No legitimate standardized test, such as NAEP or the College Board, allows school personnel to see the test questions in advance. No legitimate test uses the same exact test year after year. In addition, legitimate standardized tests only allow 50 percent of the students to test "above average." Publishers and local school authorities claimed the scores on "Lake Woebegone" tests were improving because the schools were improving. However, the actual process under way was increasingly efficient revelation to students, before their test, of the questions that would be on their test.

I decided to survey all 50 states to see if any states were testing below the publisher's "national norm." Friends for Education had not yet obtained any outside funding so I, my nurse, lab technician, and X-ray technician called and wrote letters to state education departments requesting test information. After obtaining results from more than 3500 school districts, we concluded that 95% of American school districts, and all 50 states were claiming that their local schools were above the national average on commercial achievement tests. Our study showed that some of the poorest, most desperate school districts in the nation are able to pacify the press, parents, and elected officials by testing "above the national average" on one of these sham commercial achievement tests.

The Effects on American Schools

It is important to note that the tests that give us the "Nation at Risk" message--the National Assessment of Educational Progress, the College Entrance Examinations, the International comparisons of student achievement--are not the tests American school officials use to assess local school achievement. Instead, within the last twenty years, American school board members have become dependent on one of five commercial achievement tests to measure local school progress: the *California Achievement Test*, the *Stanford Achievement Test*, the *Metropolitan Achievement Test*, the *Comprehensive Test of Basic Skills*, and *the Iowa Test of Basic Skills*. In the last 15 years, these five tests have become the principal local yardsticks, the local internal report cards of American public education.

Just as the National Assessment of Educational Progress (NAEP), the College Board, and the Armed Services Vocational Aptitude Battery are used by federal officials to measure America's educational progress, commercial achievement tests are used by local officials, parents, and the press to measure local school's progress. However, commercial achievement test publishers have not taken any of the simple security precautions with their product that NAEP, the College Board, or the Armed Services routinely takes with their tests.

Commercial test publishers even sell test preparation materials which contain review questions taken directly from currently used commercial tests. For example, the CTB/McGraw-Hill's CAT Learning Materials unethically preps students on a California Achievement Test question by telling students how to change a thermometer reading by 10 degrees. One of the questions on the most recent edition of the California Achievement Test asks students to indicate a thermometer reading that is 10 degrees higher than the one pictured.

Current testing practices victimizes school teachers as well as children. Teachers around the country have complained bitterly to me about the extent of unethical testing practices in our schools. Many teachers were concerned that if they didn't cheat, they would look bad compared to the teachers who did. All the teachers complained that cheating is encouraged by their school administrators in order to make the school's achievement scores look good.

Twenty years ago commercial achievement tests were mainly used for instructional purposes. Teachers used them to determine which students were behind and if the class needed more work in one subject than another. Class scores, school scores, district, and state scores were either not compiled or not made public. The tests were used to help children, not to evaluate educators.

However, that changed when state legislatures started insisting on accountability. Almost overnight, the tests were asked to serve an accountability purpose instead of just an instructional one. They have since become the principal local yardsticks of American educational progress. It seems unlikely that commercial achievement tests will ever again be solely instructional aids. Therefore, publishers need to modify the tests to serve their present function.

The glowing press releases, glossy student achievement brochures, "good news" parent report forms, and optimistic official "accountability" reports put out by American school officials are testimony to the fact that public educators themselves now use commercial achievement tests to measure school quality. And, for the last 15 years, American educators have found it easier to improve test scores than to improve public schools.

State legislatures and school boards need accurate measurements of local achievement. Local officials can not operate blindly, they need to know what children know and when they know it. How can local officials reform American schools when their principle yardsticks tell them they already have?

Recommendations to the Subcommittee

I endorse the recommendations of the National Commission on Testing and Public Policy and suggest you establish a "Truth in Testing"

agency to oversee the development, norming, marketing, administration, and reporting of standardized achievement tests. However, I believe such an agency should limit itself to simply protecting the American consumer against fraudulent testing, such as testing that allows all schools to be "above the national average," or tests that are administered without basic security procedures. I do not think that a "Truth in Testing" Agency should attempt to dictate testing policy to state decision makers. That is, the decision to test, when to test, and what to do with the resulting scores should continue to be the state's decision. The agency should only be charged with making sure that such testing is honest.

Second, I suggest that you direct the Federal Trade Commission to investigate commercial test publishers. Our attorney feels commercial test publishers are presently violating current FTC regulations. I include a copy of our attorney's opinion on the matter.

Third, I suggest that you require The United States Department of Education to immediately request that commercial publishers of standardized achievement tests voluntarily comply with the following set of guidelines. These guidelines are designed to assure that the selection, use, and reporting of commercial achievement tests by America's public schools will not misrepresent achievement gains, leave false impressions of relative achievement, or otherwise deceive the American public.

1. Publishers of any group administered achievement test shall take steps to ensure that only one-half of students can test above the "national norm" on their tests. Specifically, publishers should only sell current annual norms derived from a nationally representative sample of students that use their test. This would require that publishers accept responsibility for their norm's accuracy by compiling a current annual norm from a nationally representative sample of students that use their test, and that they do this annually.

2. Publishers should discourage educators from becoming familiar with test questions. Some of the publisher's test procedure recommendations encourage teachers to become familiar with test content in a manner that invalidates the inferences consumers naturally make about the overall domain of achievement.

3. Test publishers should instruct users and consumers on the need for adequate test security, and should clearly state those security precautions in their test administration manuals. Specifically, commercial test publishers should sell tests with seals on them, and with instructions printed on the test that clearly forbid teachers from reading the test in advance of administration. Publishers should also recommend that educators deliver tests to the school shortly before testing, that tests should be given to teachers on

the day of testing, and that outside test proctors be used whenever possible.

4. Publishers should only sell norms tables that accurately reflect that percentage of special education and bilingual students that are currently tested by the public schools.

5. Publishers should keep test content secure, and not allow the questions on currently used commercial tests to be used as "review" questions in test preparation materials.

Thank you for holding this hearing, and for requesting my testimony. If the committee is interested in investigating the extent of cheating by American school officials, or the effect that fraudulent testing programs have on teacher's morale, I would attempt to supply your staff with names of teachers willing to testify.

Sincerely,

John Jacob Cannell

UNITED STATES DEPARTMENT OF EDUCATION

OFFICE OF THE ASSISTANT SECRETARY
FOR EDUCATIONAL RESEARCH AND IMPROVEMENT

JUL 9 1990

Honorable Augustus F. Hawkins
House of Representatives
Washington, DC 20515

Dear Mr. Hawkins:

I recently received a copy of written testimony from Mr. Walter E. Faithorn, Jr. prepared for your hearing on testing, assessment and evaluation held June 7, 1990.

The testimony references a meeting that was held at the U.S. Department of Education at the request of Mr. Faithorn on June 4, 1990. I believe that a portion of this statement before your Committee does not accurately reflect what was said by staff of my office at the meeting. For that reason I want to correct the record pending before your Committee.

The issue involves "allegations of cheating, fraud, and deceit" in administration of standardized tests. According to the testimony, Mr. Faithorn reported that Department staff told him:

> "that not the Congress, nor the States, nor the local school boards...want the U.S. Department of Education messing around in matters of this sort--telling them what they are doing wrong, how this State compares to that State, or this school district compares to that, etc."

In fact, this view was not expressed at the meeting. Instead, my staff described the process by which a Federal agency would acquire a regulatory role in a matter such as administration of standardized tests and pointed out that the Department of Education had no such function. They also pointed out, in agreement with Mr. Faithorn, that issues of norming and test security are very important to the Federal government. This is why we replicated Dr. Cannell's first study and have asked members of the committee in charge of the Code of Fair Testing Practice in Education to consider issues of test security in the future. In addition, Department staff advised Mr. Faithorn that the U.S. Department of Education was also actively working on other, and possibly larger, issues in testing and assessment: strengthening the state-of-the-art in testing; articulating the relationship between testing, instruction and curriculum; making

Page 2--Honorable Augustus F. Hawkins

tests more "authentic" measures of what students are capable of performing; and improving dissemination of information about effective practices in testing.

Another point made by Mr. Faithorn was his concern "about a possible reduction in the rigor with which test security will be practiced" in the National Assessment of Educational Progress (NAEP). This matter was not discussed at Mr. Faithorn's meeting with my staff and I am not aware of the reason this statement was made. However, it is incorrect and would be contrary to policies and practices under which the Department carries out the National Assessment. We have made a special effort to incorporate procedures into NAEP that maintain test security. We have a strict item release policy, we maintain the confidentiality of all students and schools that participate in the assessment, and we monitor half of the schools in the Trial State Assessment and no school knows it will be monitored until the day of the assessment. We have given some consideration to the possibility of monitoring fewer sites, although of course still unannounced. But any decision along these lines would follow a careful evaluation of actual experience in 1990. At present we are inclined not to make such a reduction.

Another procedure to assure rigor and test security is that all test booklets are wrapped in plastic and are not opened until the day of the assessment and all materials are quickly collected and returned to the NAEP contractor immediately after the testing is completed. In sum, the Department would not take any action that would reduce test security or reduce confidence in the validity of NAEP results. In fact, we are continually working on ways to improve them.

You have also invited me to prepare a statement for the hearing record and I will do that separately within a few days.

Sincerely,

Christopher T. Cross
Assistant Secretary

cc: Walter E. Faithorn, Jr.

U.S. DEPARTMENT OF EDUCATION
OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT

NATIONAL CENTER FOR EDUCATION STATISTICS

JUL 12 1990

Honorable Augustus F. Hawkins
Chairman
Committee on Education and Labor
House of Representatives
Washington, DC 20515

Dear Mr. Chairman:

I appreciate your invitation of June 6 to provide a statement for the Committee record on the subject of your hearing dealing with testing, assessment, and evaluation. These areas of education measurement are of central importance to the National Center for Education Statistics (NCES) because of the growing interest in assessing student performance and because NCES uses tests for a number of its data collection activities. While we do not face the challenges or needs of schools and districts in relating institutional goals, curricula, and instructional materials and methods to testing, we do draw on the available expertise and are influenced by the same debates about testing in which schools, districts, States, researchers, policymakers, and the public are currently engaged.

Let me respond to your questions in turn:

1. What are appropriate measures to assess learning in our schools?

   There is no single test or test format that is appropriate for measuring all learning in our schools. The measure of progress toward learning of a specific curriculum requires a criterion-referenced test (a test that measures how much has been learned from a well-defined domain of content skills). This type of test is used in most State and local testing programs. Measuring progress toward broadly defined objectives and making comparisons among groups require a norm-referenced test (a test that compares a student's skills to those of other students), such as the tests provided by commercial testing programs and the college entrance examinations.

   Currently available criterion-referenced and norm-referenced tests are not appropriate as exclusive indicators of the comparative progress of different educational systems. For example, the reports by John Cannell that were described to

WASHINGTON D.C. 20208

your Committee clearly indicate that norm-referenced tests can give misleading results when used for this purpose, because administrative practices are not uniform. The National Assessment of Educational Progress (NAEP) is the only currently available assessment that is specifically designed to measure trends in the progress of education systems and make comparisons among States.

2. How can testing and assessment programs at the Federal, State, and local levels be integrated and interrelated?

It is important to continue separate testing programs at the national, State, and local levels because each type of program is specifically designed to serve a different function, as noted above. Local testing programs should evaluate student learning of the specific local curriculum, and they are used to diagnose individual student strengths and weaknesses, and to assist classroom teachers with instruction. Local testing programs cannot be aggregated to evaluate progress across States or the Nation because they cover different content in different grades, at different times and under varying procedures. NAEP, on the other hand, can be used to evaluate progress of the States and the Nation against a standard set by consensus, but is not appropriate for evaluation of school districts, schools, or students because its content is not specifically aligned with curricula studied in each district and classroom.

However, various assessment programs can be articulated or connected through different "linking" mechanisms. In the NAEP trial State Assessment, NCES is encouraging and providing technical assistance that will make it possible for States to link their State testing programs to NAEP. Once this is done, a State can provide a "NAEP equivalent" score to all of its students who were tested in the same subject in the same grade (e.g., for all math students in the eighth grade) in the State. In this way, a student's, school's, or district's score would be more valuable because it could be compared with the benchmark national scale available in NAEP.

3. How can we minimize the possible adverse effects of testing?

The issue of adverse impact is most relevant to local and State testing programs that are used to make decisions (such as promotion, graduation, and program placement) about individual students, and to such testing programs as the Scholastic Aptitude Test (SAT). Such tests, if they are biased or otherwise unfair, may deny students educational and employment opportunities. The assessments administered by NCES are not used to make decisions about individual students, but instead are used to inform policymakers and

educators about the progress of education in the Nation, regions of the country, States, or various groups of students such as minority populations. To maximize the reliability of assessments used to understand the relative achievement levels of these groups, NCES undertakes vigorous examinations for bias and other forms unreliability, prior to test administration, for every item in assessments NCES conducts.

The best long term approach to minimize the potential for adverse impact is to encourage the testing profession to continue developing professional standards. The two major documents that deal with this issue are the Joint Technical Standards and the Code of Fair Testing Practice in Education (both sponsored by the American Psychological Association, the American Educational Research Association, the National Council of Measurement in Education, and other national groups). These documents have been widely endorsed by testing programs throughout the country, and represent the standards to which the profession has agreed to make itself accountable. This approach to building and maintaining standards should be refined and continued. The Center requires its contractors to follow the guidelines in these documents for tests conducted for NCES.

4. How can comprehensive assessment systems be developed at the national, State, and local levels that will focus upon student progress and school improvement for all children?

As I mentioned above in reference to the point about integration and interrelation of Federal, State, and local testing and assessment programs, assessment has a unique role to play at each level. The National Assessment program is currently a method for obtaining information on how children in American schools at grades four, eight, and twelve perform in selected subject areas. It is intended to serve as an indicator of what American students as a whole know and can do. The new Trial State Assessment collects consistent and uniform information about student performance across all States. This program will provide a way to understand the relative standing of States in terms of student achievement in given subjects, such as math and reading, and the relative strengths and weaknesses within these broad subject areas, such as the relative performance in algebra and reasoning skills. It will not provide information at the district, school, or student level, nor provide information about what changes ought to be made.

State assessment programs, in contrast, focus specifically on State level curricula and allow States to evaluate how well their districts and schools are doing in achieving the goals of those curricula. District, school, and individual

student testing programs, in conjunction with State assessments, allow local superintendents, curriculum specialists, principals and teachers to evaluate the performance of individual students and to diagnose their specific strengths and weaknesses at a detailed level. For example, such testing programs may provide information on which subtopics within the local curriculum each student has learned (e.g., in reading comprehension, whether a student can identify specific information, identify the main idea, or apply that information to a new problem).

These separate components form a whole assessment system, when each is implemented at its appropriate level--national, State, or local. Such a system would provide the information for educators and parents to know about and gain insight into student progress and school improvement for all children. Each level of assessment provides specific and unique types of information to achieve this objective.

5. What is the appropriate Federal role for improving testing and assessment at the national, State, and local levels?

In his separate reply to your June 6 letter, Assistant Secretary Christopher T. Cross addresses the overall issue of the Federal role in testing and assessment. My comments deal with the specific activities of the National Center for Education Statistics.

The Center, as I noted above, administers many tests in connection with its mandate from Congress to gather and report data on the condition and progress of education. In addition to the National Assessment, tests are used in our longitudinal studies, international achievement comparisons, and adult literacy assessments. Other areas, such as school readiness and college level achievement, could be added in future data collections.

NCES makes use of the strongest and most diverse advice it can find in developing these tests, but we are now planning to search more aggressively for approaches to testing that will make our data more reliable and valid in the future. We are exploring the possibility of supporting research and developmental work needed to improve the state of the art for large scale national and international assessments. Some of the areas include: incorporating recent findings in cognitive psychology into educational assessment instruments, using computer technology to assess the learning strategies of students, improving psychometric procedures for "authentic" performance test items, and improving methods of measuring "opportunity to learn" in international assessments.

Page 5 - Honorable Augustus F. Hawkins

Even though our purpose in these activities is to make it possible for NCES to report more reliable, valid, and complete statistics on education, this new knowledge would be of direct use to States and other sponsors of large testing programs as well. Thus, NCES would be able to provide technical assistance to other education data collectors. In addition, NCES will be supplementing the activities of the National Cooperative Education Statistics System so that States and local districts can increase and improve their efforts to monitor progress toward the President's and the Nation's Governors' national education goals. This activity will lead to improved data and indicators that would be tailored to local conditions and needs.

Thank you for providing this opportunity to comment on these important testing, assessment, and evaluation issues you have addressed in your Committee. If I or members of the NCES staff can provide further assistance, please let us know.

Sincerely,

Emerson J. Elliott
Acting Commissioner

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
08541

GREGORY R. ANRIG
PRESIDENT

July 19, 1990

The Honorable Augustus F. Hawkins
Chairman
Committee on Education and Labor
U. S. House of Representatives
B-346C Rayburn House Office Building
Washington, DC 20515

Dear Representative Hawkins:

I want to take this opportunity to tell you how much I admire the outstanding leadership you have provided in the House of Representatives over the past 28 years to further the cause of equal employment opportunity and quality education. It has been a pleasure working with you and your staff on the important educational issues during this period of time.

I understand that the record is still open from the June 7th hearing on educational testing and assessment. I would like to respond to your request for comments on several important issues which, not surprisingly, are priority concerns for us at ETS.

The issue of appropriate measures to assess learning is of central concern to us as it is to you. There are many different approaches to assessing important aspects of student learning. Multiple choice tests are most widely used for assessment of learning in situations in which large-scale and low-cost assessment is needed. However, even projects the scale of the National Assessment of Educational Progress (NAEP) include non-multiple-choice portions, with 30% of its recent assessments calling for performance responses by students.

Today there is a great concern for so-called authentic assessment or performance testing. Such assessment may be as important for the impact it has on the educational system as for the types of learning it assesses. At ETS, we have made some exciting advances in large-scale performance assessment ranging from the National Assessment to scoring hundreds of thousands of student essays each year for the College Board's Advanced Placement Program to performance testing in licensing programs of several types. Such assessment is more expensive than multiple choice testing as well as requiring more student time and judgmental scoring, but it is practically feasible.

In addition, we are very excited about the role assessment can play in improving learning at the classroom level. We have several projects in which assessment is designed specifically for the purpose of improving student learning. In an experimental middle school science

program, the teacher uses complex integrative tasks as both instruction and assessment. In Arts PROPEL, a project in collaboration with the Pittsburgh Public Schools and Project Zero at Harvard, teachers and students use the assessment of student portfolios of art and writing as part of the learning and instruction process. In these and other activities, we are putting assessment to use directly for student learning. The resulting assessment is quite different in nature than assessment designed for judgments of student learning independent of the classroom learning context.

These differences in assessment we are seeing at different levels and for different purposes relate directly to the issue you raised of how to integrate assessment programs at various levels. Clearly, there needs to be more connection between what is good for the classroom and what is used for large-scale evaluation and accountability. We suspect that the route of NAEP with a combination of economical and efficient measures supplemented by a substantial portion of performance-type measures is a useful approach for accountability testing at the Federal and state levels. In the long term, however, as we learn more about the complex forms of assessment that now seem feasible only at the classroom level, it may be possible to accomplish a more thorough integration.

In a recent article, I described two other important assessment concerns, excessive testing in this country and efforts to insure fairness. I am enclosing that article as well as NAEP background information for your reference.

I understand that the Committee may hold future hearings on the subject of educational testing. I would like to offer whatever information or assistance ETS can provide to help you in your examination and deliberations on the important issues you are addressing. Thank you for this opportunity to contribute to the record of the recent hearing.

I wish you good health and much happiness in your retirement. We will miss you behind the center seat at hearings and remain grateful for your tireless efforts on behalf of equal employment and educational opportunity in America.

Sincerely,

Gregory R. Anrig

Attachments:
The NAEP Guide
Brochure on Innovations in NAEP
Standardized Testing - Now and in the Future, by Gregory R. Anrig - Article from the Alumni Review of the Harvard Graduate School of Education, Spring 1990

First two enclosures have been maintained in Subcommittee files.

STANDARDIZED TESTING - NOW AND IN THE FUTURE

Gregory R. Anrig
President
Educational Testing Service
January 1990

What an extraordinary time for standardized testing in American education. In September, the President of the United States and the nation's governors meet in a landmark "education summit" and jointly call for national performance goals for education and a means to measure progress towards these goals. In the same month, the Annual Gallup Poll on Public Attitudes Toward the Public Schools is released. Of those polled, 70 percent favor national achievement standards and goals, and 77 percent favor the use of standardized national testing programs to measure the academic achievement of students.

But this is just the tip of an iceberg! The 1980s have seen an explosion of standardized testing in education. Forty-four states now require some form of minimum competency tests, 35 of them requiring the use of state-developed or state-selected tests with state-prescribed performance standards. Twenty-one states have testing requirements for high school graduation. Where only a handful of states had testing requirements for the initial certification of teachers in 1980, this year 45 states have such testing requirements. And the National Board for Professional Teaching Standards is in the process of developing new assessments to recognize advanced teaching ability of experienced teachers.

Some Personal Perspectives on Standardized Tests

I was a consumer of tests before coming to ETS in 1981. I had a healthy skepticism of standardized tests as a teacher, principal, superintendent, and state education commissioner. It may surprise you that, after eight years as ETS President, I still have a healthy skepticism of standardized tests. I am an educator, first and foremost, and I judge tests and other information on the basis of how much they help learning and the improvement of education.

Standardized tests do provide educationally useful information -- when properly used, properly interpreted, and used in conjunction with other information before making decisions. They provide a useful "check and balance" on other information precisely because they ARE standardized in content and administration. For those who fault this standardization, consider the alternatives! I remember well the fatigue, stress, and uncertainty that accompanied the homemade tests I developed and graded at night as a teacher of junior high school social studies.

It currently is de rigueur to criticize standardized tests in general and the poor old multiple choice question in particular (a format, by the way, that reliably measures much more than its critics say and at a lower cost to the taxpayer or parent). Thanks to C-SPAN, I observed the testimony of educators before the National Governors' Association Task Force on Education. One after another decried the use of standardized tests to judge the results of education.

Prepared for the Spring 1990 edition of the Harvard Graduate School of Education Alumni Bulletin.

I strongly believe this is an ill-advised position for educators. The public and their elected officials want to know what students know and are able to do. They have a right to know this and educators had better find a way to be responsive. We need to remind ourselves that the education reform movement of the 1980s got started because of public concern that children were not learning enough in schools or even as well as they used to. This concern was justified then, and it still is.

I believe that some standardized and economically feasible way of assessing what is learned by students will be required of educators as an outcome of the historic new commitment to national performance goals. Once this is accepted, then we can focus realistically on the cost, time and content trade-off issues related to such standardized assessments.

Three Key Issues of Standardized Testing

Although I applaud the new commitment to goals, I am concerned that it may lead to an unnecessary proliferation of testing in American education. I spend much of my time outside of ETS telling people that too much time and money are being spent now on accountability testing. We test too little too much. It is like pulling up a carrot to see if it is growing. Can they read? Can they read? Can they read? We can get a good answer to that question without testing every child several times every year.

The National Assessment of Educational Progress (NAEP) has demonstrated methodologies that can avoid the overuse of accountability testing by states. NAEP assesses samples of knowledge and samples of students accurately and reliably. ETS has been proud to administer NAEP since 1983 and I believe it is becoming a creditable "report card" for America's schools. One sign of that credibility is the fact that 37 states have signed up for the new state assessments authorized by Congress in 1988. My hope is that this new resource will help states to reduce the time and money already being devoted to state accountability testing.

In addition to urging people not to test so much, I try to counsel them about the importance of keeping tests in perspective and using them properly only for the purposes for which they are designed. The SAT, for instance, is a college admissions test and never was designed to measure the overall quality of American schools. Yet Secretaries of Education and the media continue to use it improperly for this purpose. The National Collegiate Athletic Association decided to use SAT and ACT scores -- improperly, in my judgment -- as eligibility criteria for freshman athletics. Arkansas and Texas sought to use scores from the NTE to determine whether in-service teachers could continue to teach. In each of these cases, ETS has publicly opposed such improper test use (even refusing NTE services to Arkansas and Texas) and has offered pro bono assistance to develop proper alternatives.
In addition to these efforts to promote proper test use, ETS last year joined with five other major test publishers and publicly adopted a Code of Fair Testing Practices in Education. I believe that testing organizations like ETS have a public responsibility not only to develop the best tests possible but also to be strong advocates for the proper use of these tests.

For me, the most troubling issue regarding standardized testing in American education is test bias. Tests are made by human beings and therefore certainly can be biased. Organizations that develop tests have a fundamental responsibility to guard against test bias. I am proud to say that probably no organization works harder to assure fairness in testing than does ETS. Every question on every form of every test that ETS develops must go through a mandatory sensitivity review. Specially trained staff search for any indication of bias, using structured guidelines and procedures. Committees of external experts in each discipline scrutinize test items and performance statistics. Internal and external audit teams annually review adherence of each testing program to ETS's Standards for Quality and Fairness. ETS and its clients regularly conduct research on test bias and publish the findings and data for scrutiny by independent researchers. Those who use ETS-developed tests are given guidelines and training on their proper use and interpretation. In addition, a new statistical procedure was introduced in 1987 and now is applied to every ETS-developed test. Called Differential Item Functioning (DIF), it provides a means to analyze the performance of students of like ability on each test question, based on the student's race, sex, and ethnicity, before the question is used for scoring. A major step ahead in guarding against test bias, other testing organizations are following ETS's lead and are using DIF for their tests as well.

It is essential that there be continuing scrutiny, debate, research and critical analysis regarding test bias. Those who develop and use tests or are affected by them should be part of this ongoing process. And what is learned should be used to change test development practices. I am troubled, however, by the trend of some critics of testing and some of the media to define "bias" simply as meaning any difference in test results by race, sex, or ethnicity.

Unequal educational opportunity regrettably is still a reality in American education. It is essential that the public spotlight continue to focus on these unequal opportunities until they are corrected. Tests are an invaluable resource for demonstrating the profound effects of such inequalities. In recent years, a number of nationally standardized tests have begun to report improved academic performance of minorities and women as their educational opportunities have improved. It is shortsighted advocacy to call for moratoriums on the best vehicle for promoting public action against educational inequality.

There is a moral and educational imperative to guard against bias in standardized tests. ETS and I fully respect and accept our responsibility for this imperative. But there also is a moral and educational imperative to determine fairly and report clearly any differences in academic achievement that exist among students, regardless of race, sex, or ethnicity. To call this bias is a serious mistake.

The Future: New Kinds of Standardized Tests for New Kinds of Purposes

I came to ETS because I believed it had a unique capacity to help public education shape clearer and higher expectations for learning and to create a new generation of standardized assessments to usefully measure this learning. In 1987, the ETS Board of Trustees approved a five-year plan to achieve these aspirations and committed a major share of ETS's financial resources to fund the effort. We are midway in this undertaking and already are seeing what this new generation of educational assessments can be.

Some of these assessments will be performance based. ETS and Harvard Professor Howard Gardner are working with teachers in the Pittsburgh Public Schools on portfolio assessment of student work in art, music, and creative writing. Here teachers are being trained to assess student work products at the draft stage in order to guide students to the next level of accomplishment. In another field, ETS researchers are developing a computer-based science program for middle school students. Students will solve problems and conduct experiments, receiving continuous feedback on how they are progressing. A successor to the NTE is under development that will involve three stages for teacher licensure. The third stage will be services to promote state policies for systematic assessment of actual teaching performance in the classroom as one part of initial licensing requirements for beginning teachers.

A second characteristic of these new assessments will be that they increasingly will be instructionally-based. Most current standardized tests are not very useful for the classroom teacher. Some of the new assessments will be designed specifically for the teacher. A new publication called ALGEBRIDGE will be released in 1990-91. It is aimed at introducing middle school students to algebra. Field tested with teachers and students in six urban school districts, it provides assessment information to students and teachers as they tackle basic concepts essential to an understanding of algebra. The purpose of ALGEBRIDGE is to encourage more urban students to elect algebra in ninth grade as part of a concerted effort to promote their access to college. To improve critical thinking skills, a computer-based program is being developed in several New Jersey and Massachusetts schools as a part of middle school language arts programs. Again, assessment will be aimed at giving immediate feedback to students and teachers.

A third characteristic of some of these new assessments will be the use of technology and new forms of adaptive measurement. In a project for the National Council of Architectural Registration Boards, computer-based certification examinations will involve actually doing design projects and calling upon the standard references found in an architect's office. The GRE Board has just launched a research and development project to computerize the Graduate Record Examinations. The computer will simulate actual tasks that graduate students regularly are called upon to do, such as reference searches, and will automatically move students to tasks at higher or lower levels of difficulty depending on their performance. ETS also is developing new adult literacy assessments that are designed to aid employers and job training centers in raising literacy skills employees need for the changing workplace.

These are very different kinds of assessments from the current standardized tests available to American education. As can be seen, their purpose is not accountability. Their primary purpose will be to draw upon advances in technology, cognitive science, and measurement sciences to provide information that is useful to learners and teachers.

We are at the threshold of dramatic changes in standardized educational testing. These changes are not limited to ETS's efforts. They are going on elsewhere as well. At ETS, the focus will be on new assessments that promote the improvement of learning and of educational opportunities. These changes are not dreams. They are initiatives already begun that will yield significant results in the 1990s. This is indeed an extraordinary time for standardized testing in American education.

FairTest
National Center for Fair & Open Testing

June 21, 1990

Augustus F. Hawkins
Chairman
Committee on Education and Labor
U.S. House of Representatives
B-346C Rayburn House Office Building
Washington, D.C. 20515

Dear Chairman Hawkins and Members of the Committee:

The National Center for Fair & Open Testing (FairTest) is pleased to respond to the Committee's invitation to offer testimony in writing on the subject of educational testing and assessment.

Before entering my discussion, let me summarize FairTest's two broad recommendations:

- The federal government must stop mandating educationally harmful forms of testing and assessment.

- The federal government has a potentially valuable role to play in supporting district, state and federal government development of educationally helpful methods of assessment.

FairTest is the nation's only organization solely dedicated to making testing and assessment fair, open and educationally relevant. FairTest has found, however, that because of lack of accountability by the testing industry, conceptual flaws in the design of most tests, and the misuse and overuse of tests, much of the testing that is done today is educationally destructive.[1]

Testing exerts its harmful effects in three basic ways:

First, the most prominent role of testing has been to exclude racial and ethnic minorities, women and the poor. Indeed, the ability of tests to sort people by these categories was a major reason for the popularity of early standardized tests. While there have been instances where testing has opened opportunities, since its early days testing has served primarily as a gatekeeper.

Second, as standardized, criterion- and norm-referenced multiple-choice tests emerged as the most important part of school accountability programs in the 1980's, they came to exert a powerful, often controlling influence on curriculum and instruction in the schools. As

342 Broadway, Cambridge, Mass. 02139 (617) 864-4810 FAX (617) 497-2224

many studies have indicated, they exert the most influence on programs and classrooms populated by students who score low on tests, because it is those programs which try hardest to increase test scores. These programs are disproportionately filled with low-income and minority-group children. As a result of the focus on testing, these students read less, write less, do fewer projects, do not use their higher order thinking abilities in school, ultimately do not become proficient students, and frequently drop out. Testing encourages, reinforces and justifies all these harmful trends.

Third, the very nature of multiple-choice testing presents incorrect ideas of how people learn. While cognitive and developmental psychology have conclusively shown that humans learn through active engagement with the world, multiple-choice testing is rooted in outmoded behaviorist psychology that views learning as the passive accumulation of isolated bits of information. Even in learning "basic skills," students use higher order thinking processes, but the tests artificially and incorrectly separate basic from higher order.[2] As the tests have come to control curriculum, they have encouraged a completely incorrect approach to instruction in the effort to raise test scores in the short run.

Taken together, these three points paint a sad picture: too many students are tracked using tests and placed in "dummied-down" programs where they are not challenged or stimulated and fail to make adequate educational progress. These students are disproportionately low income and children of color. The evidence leads to one essential conclusion: **our nation's efforts to construct schools worthy of our children will fail so long as standardized, multiple-choice testing remains the coin of the educational realm.**

However, in our criticism of testing we must not forget two important objectives that testing promised - but failed - to meet: to provide assessment and evaluation information that teachers and administrators could use to improve instruction, and to provide information on student and program performance for accountability purposes. Both these goals must be met, but they must be met in a manner that does not end up sabotaging the fundamental goal of improving public education, as the tests have done.

What then can be done and what is the federal role? FairTest makes the following factual observations and from them offers recommendations.

In state after state across our country, departments of education are working to develop performance-based assessments. This type of assessment asks students to work on real tasks, thereby directly demonstrating knowledge and capabilities, rather than fill in bubbles on multiple-choice questions. This process not only provides valuable information about achievement, it also fosters instruction that encourages thinking, exploration, reflection, cooperative learning, and, through them, the acquisition of and ability to use various skills and factual information.

Plans by states to develop and use performance-based assessments are expanding rapidly. At the June 1990 Education Commission of the States (ECS) Conference on

Assessment, a number of states that have mandates to develop and use the new assessments agreed to form a consortium. The states will share resources in developing and analyzing portfolios, open-ended test items and other forms of performance-based evaluation. This emerging consortium, to be co-ordinated by ECS, is only one of several being developed.

Additionally, many states are actively engaged in transforming their state assessment systems. Among these states are California, Connecticut, Vermont, Arizona and Kentucky; many more are investigating how to begin this process. Also, many districts are actively engaged in efforts to transform their assessment systems as part of changing to school-based management and adopting new models of curriculum and instruction.

Performance-based assessment is still emerging, so much remains to be learned and many problems must be solved. Research and experimentation, however, indicate in outline form what a performance-based system can look like.

At the classroom level, the essential tool is the portfolio. Portfolios are not simply a place to dump all a student's papers. Rather, they are tools for reflection and evaluation. They enable teachers and students, as well as parents and administrators, to see progress students make toward agreed-upon educational objectives. They facilitate diagnosis of strengths and weaknesses, indicate the student's individual work that should be done, and demonstrate the achievement. They also presume that something worthwhile is happening in the classroom; to fill a portfolio with ditto sheets and answers to multiple-choice questions taken from basal readers is simply a waste of time.

At the state level, there are two essential assessment tools. One is evaluation of portfolio work. Vermont, for example, will look at a sample of portfolios in every school in the state in grades four and eight. This will enable the state to report on student achievement, note progress and problems, and make recommendations to both schools and individual teachers. Because teachers will be trained as portfolio evaluators, a great deal of staff development in new forms of instruction and evaluation can take place. It is important to note that portfolios can be assessed in ways that provide aggregatable, quantitative data.

The second essential tool for states, and even districts, is the performance-based test. Such tests are easiest to conceptualize in the arts: one assesses a student's ability to play an instrument by listening to a recital. Both artistic and athletic competition, such as gymnastics, have a long history of rating performances with high levels of reliability among the raters.

Performance-based tests can take a variety of forms. On the one end, they can be "best pieces" from portfolios. That is, an important student project, such as a piece of scientific research or historical investigation, can be assessed as a test. These are tests that not only are not secret, but that must be open and serve instructional as well as evaluative purposes. At the other end are tests in the more traditional sense, only with items that force students to solve ill-structured and open-ended problems in which they first have to decide what the problem is, then offer a solution which they can explain and justify. As with

portfolios, performance-based tests provide a basis for staff development and changing curriculum. They, too, can be used in ways that provide aggregatable, quantitative data.

We must point out here that federally-mandated testing programs, particularly Chapter I, are perceived as a major obstacle to assessment reform by educational leaders at the state and district levels. So long as the multiple-choice measures are the essential tools to evaluate student and program progress, they will control curriculum and instruction and prevent districts and states from changing assessment and instruction to meet the needs of the students.

These observations lead to three recommendations:

- The federal government should support research and development at both the district and state levels in constructing, introducing and evaluating a variety of performance-based assessments, and support staff development to take advantage of needed curricular, administrative and assessment reforms.

- The federal government can help develop methods of evaluating, quantifying and aggregating educational information from performance-based assessments.

- The federal government must stop requiring forms of assessment that are educationally harmful. In particular, Chapter I testing requirements must change not later than in the 1992 re-authorization, and the National Assessment of Educational Progress testing must not be allowed to control national education with multiple-choice testing.

Experience over the past decade has shown that over-emphasis on one form of assessment, the multiple-choice test (both norm- and criterion-referenced), has harmed our nation's ability to make needed changes in curriculum and instruction. While teaching to even a modestly adequate performance-based item would be superior in many ways to teaching to any multiple-choice test, the danger of educational and evaluative corruption remains.

For example, in woodworking a performance-based curriculum and assessment could have a student construct a chest of drawers. Properly used, teaching to this task would have students explore many alternatives in construction, choose one and defend the choice, then actually make it. Incorrectly used, the teacher would insist on a narrow range of construction possibilities (for example, only one kind of joint), teach only that narrow range (indeed, repetitiously drill on the one joint), in order for the students to do well on just the one project. The result may be high scores on the chest of drawers, but the students would not have learned enough to solve any other problems. i.e. make other types of cabinets requiring other types of joints. Thus, both curriculum and instruction and assessment would be corrupted: the students would not learn broadly, and to the extent the work sample was

supposed to represent a broader domain of learning cabinetry, the results would be misleading and invalid.

The problem is that when heavy pressure comes down on administrators and teachers to ensure that students perform well on narrowly-defined tasks, even if they are performance-based, they will tend teach to the test in a narrow way and to the exclusion of other, needed areas. The tendency is also to over-emphasize what the teacher wants (regurgitation) to the exclusion of student exploration (guided, active learning). Both instruction and assessment are thereby damaged, and both students and society suffer. The question is, how can the federal government, the states or the districts use testing for accountability purposes without sabotaging the instructional process and narrowing the curriculum?

At this point, FairTest believes there are several parts to an answer.

First, the primary goal of assessment must be enhancing the quality of instruction. Making portfolios the basis of assessment, with various types of tests established as complements, well serves this purpose. Portfolios can then be evaluated in terms of goals too broad and complex to allow teaching in a narrow manner. The Advanced Placement art portfolio assessments conducted by Educational Testing Service are an example of this: A vast array of artwork, including a portfolio of best pieces and slides of a range of work from each student, are evaluated by teams. Many kinds of art are judged as having artistic merit; what is essential is the student's display of implementing her or his vision, of having an artistic voice he or she can put into effect using artistic techniques. There is thus neither need nor ability to teach narrowly to a narrow test. At the end however, it is possible to assign a number, or set of numbers, to each portfolio, on the basis of agreed-upon criteria, and these numbers can be the basis for quantifiable, aggregatable data.

Second, where testing external to the classroom exists, it should be done on a sampling basis and there must be sufficient items so that it becomes impossible to teach to any one or few items. This will require developing a large number of good items and training evaluators to evaluate such a wide range of items. It also requires developing the capacity to equate many complex items so quantification becomes possible.

In short, variety, complexity and richness of forms of evaluation, guided by the understanding that without good activities in the classroom real learning will not take place, are the only means of dealing with the problem of corruption.

- FairTest recommends that the federal government help fund a variety of assessment activities, giving primacy to those that encourage staff development through teacher involvement and that are most useful in instruction. These may be developed by districts, the states or even the federal government, but must be focused on improving instruction first and providing aggregatable data second.

34-661 - 90 - 5

In designing and implementing new forms of assessment, many complex questions must be resolved. Three more are important enough to require consideration here.

First, removing or reducing the use of multiple-choice tests that are biased and introducing new forms of assessment that encourage thinking does not mean that the new forms will not be biased. As new assessments are introduced, it is essential that several things be done to reduce and eliminate bias. One, all students must be enabled to understand the meaning and processes of the new assessments. Two, the evaluation process constructed around portfolios must incorporate methods to detect and address teacher bias. Not only are portfolios a valuable means of helping teachers become better instructors in the subject areas, they can be valuable methods of helping teachers overcome the ignorance that underlies much biased behavior. Evaluation through portfolios, coupled with interviews and classroom observations, can provide a basis for educating most teachers and removing those who refuse to change and grow.

Second, true performance-based assessments are not likely to have much in common with multiple-choice tests because they are not likely to measure the same things.[3] As a result, complex problems may develop for longitudinal, continuity data. FairTest strongly urges that the desire of federal or state agencies to limit performance-based testing in order to preserve continuity data be subordinated to the far more critical need to introduce well-developed performance-based assessments in order to assist fundamental school reform. Controlling the new to meet needs rooted in the old that has failed is only a means to guarantee continued failure. Research on how to bridge data from the two means of assessment to continue making use of old data could be useful, but funding for such research also should be subordinate to developing and implementing performance-based assessment.

Third, issues of reliability and validity of performance-based assessments need continued investigation so as to enhance their quality. Federal funding to help such studies of new assessment programs as they are designed and introduced would be a valuable use of the federal dollar.

In summary, the federal government has a valuable role to play in changing assessment in our nation. Through well-directed funding and changing certain laws, the federal government can open up the possibility of using appropriate assessments and hasten the implementation of high-quality performance-based assessments.

To do the latter, funds must be carefully targeted. It is clear to FairTest that the most exciting and powerful developments are now happening at the state level, both within particular states and among states acting in consortia. While many districts and even individual schools and programs are actively engaged in needed assessment reform, it is at the state level that change which is both extensive and profound can best be developed. That

said, it is also clear that only when schools, teachers, administrators and parents are actively involved in the change process can reform really take hold in a comprehensive way.

Therefore, both the states and the districts are essential to the change process, but they have different roles. The role of the state is to develop and disseminate possibilities for performance-based assessment, beginning with their own assessments, and including extensive teacher education in portfolio and other assessments, as Vermont plans to do. The role of the districts is to implement forms of portfolio-based assessment as the core of instructional evaluation and to create processes of renewal that unleash the creativity and capabilities of all people working in and for schools.

The federal government can and should act to facilitate this process. In funding, it should fund at both the state and district levels, and funding at one level should require interaction with the other level. States not working with districts are apt to develop unused procedures or re-visit the failures of top-down dictates. Districts not working with states are apt to change in isolation and fail to help wider change, or to run afoul of state regulations that undermine local change.

FairTest thus urges the federal government to proceed in the direction of encouraging, through funding and changes in law and regulation, the development and implementation of performance-based assessment that builds from the classroom up and that supports an instructional process that encourages thought, reflection, activity, engagement and creativity as ends in themselves and as the best means of developing basic and more advanced academic skills.

The federal government's steps in this direction should come soon, but must not be taken too hastily. We urge the federal government to use the principles and guidelines discussed above, or similar ones emerging now from many sources including state departments of education and academic researchers. The government should carefully but expeditiously develop plans to assist fundamental change in assessment in our nation's educational systems, and thereby enable the needed changes in curriculum and instruction.

Thank you again for the opportunity to testify.

Monty Neill

Monty Neill, Ed.D.
Associate Director

Attachments:
Endnotes
Fallout From the Testing Explosion
"Standardized Testing: Harmful to Educational Health

NOTES

1. Medina, Noe and Neill, D. Monty. *Fallout From the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools* (Cambridge: FairTest, Third Edition, 1990). This report summarizes, with extensive evidence and notes, the many problems associated with the tests; it includes an annotated bibliography. A copy is appended. Portions of the report appeared in revised form in Neill, D. Monty and Medina, Noe J. "Standardized Testing: Harmful to Educational Health," *Phi Delta Kappan* (May 1989) pp. 688-697; a copy is appended.

2. Resnick, Lauren B. and Resnick, Daniel P. "Assessing the Thinking Curriculum: New Tools for Educational Reform," in B. R. Gifford and M.C. O'Connor, eds., *Future Assessments: Changing Views of Aptitude, Achievement, and Instruction* (Boston: Kluwer Academic Publishers, 1989).

3. Newmann, Fred and Archbald, Doug. *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School* (Reston, VA: National Association of Secondary School Principals, 1988), see esp. pp. 56-59; Fredericksen, Norman. "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist* (March 1984) pp. 193-202.

American Psychological Association

Advancing psychology as a science, a profession, and as a means of promoting human welfare

June 19, 1990

Committee on Education and Labor
U.S. House of Representatives
B-346C Rayburn House Office Building
Washington, D.C. 20515

In response to recent concerns over the use of standardized tests in education, the American Psychological Association (APA) is submitting this statement to the House Education and Labor Committee, following its June 7 1990 hearing on Testing in Education.

APA has historically supported scientific and policy initiatives that have improved the development and use of assessment practices and instruments. APA and its divisions have developed professional standards in these areas which have been widely accepted in legal and legislative arenas (e.g., Standards for Educational and Psychological Tests, Code of Fair Testing Practices in Education). In addition, APA has several standing and ad hoc committees which are charged with addressing critical issues in assessment (e.g., Task Force on the Prediction of Dishonesty and Theft in Employment Settings, Joint Committee on Testing Practices).

There are several areas of concern regarding the use of standardized tests in education, many of which are outlined in the report From Gatekeeper to Gateway, produced by the National Commission on Testing and Public Policy. These concerns include the amount of testing that takes place in our schools, some inappropriate uses of test score data, and the overreliance on test scores alone in making decisions about individuals. APA believes that much of the problems associated with standardized tests (in education and elsewhere) are not inherent in the tests themselves but rather are founded in their inappropriate use.

If standardized tests are properly normed and validated, they can offer information about an individual that cannot be obtained from other sources. It has always been deemed inappropriate to use test score data on an individual to the exclusion of other information. Test scores taken in conjunction with other information (e.g., grades, teacher reviews, etc.) can enhance our ability to make better and more informed decisions about an individual's educational needs and past achievement.

1200 Seventeenth Street, N.W.
Washington, DC 20036
(202) 955-7600

It is also deemed inappropriate to use test score data in a manner other than that for which the test was developed. Educational achievement tests which are designed as diagnostic tools to aid teachers in meeting an individual student's educational needs should not be used as measures of a school's educational progress. The opposite is true as well, that measures of educational effectiveness, such as the National Assessment of Educational Progress (as currently designed), should not be used for individual assessment. With this kind of proper use of standardized tests, practices such as "teaching to the test" may be curtailed.

APA's approach to such issues with testing has been one of education and training. We believe that, when properly developed and validated, standardized tests can enhance our ability to make important decisions about individuals, if they are used as they are intended.

At present, there seem to be many proposals for advancing methods of alternative assessment. APA supports ongoing scientific inquiry into the alternative assessment approaches and performance-based testing and evaluation in education. APA's concern is that alternative measures be reliable and valid. Many proposed alternatives - teacher observations, exhibitions, portfolios of student work, checklists, and open-ended questions - have not been demonstrated to have adequate reliability or validity. Historically, standardized measures were developed to correct this problem.

Additionally, proponents of alternative measures see them as correcting the problem of cultural bias in testing. In fact, many such alternatives have been demonstrated to be more susceptible to idiosyncratic beliefs or subjective judgement than traditional standardized measures. Where actual differences between groups exist, the introduction of alternative approaches may mask but will not eliminate these differences. By masking these differences compensatory strategies designed to enhance opportunites for disadvantaged groups may be lost.

As a developer of professional standards on educational and psychological testing, the American Psychological Association remains extremely interested in the quality of assessment instruments and measures that are used and developed. APA supports the scientific research into alternative assessment approaches to assure that any assessment methods used to make decisions about individuals be reliable and valid. We look forward to the Office of Technology Assessment's study in this area and hope there is a strong emphasis on examining the reliability and validity of new assessment approaches.

Sincerely,

Lewis P. Lipsitt, Ph.D.
Executive Director for Science
American Psychological Association

Wayne J. Camara, Ph.D.
Director for Scientific Affairs
American Psychological Association

cc: Committee for Psychological Tests and Assessment
Raymond D. Fowler, Ph.D., Chief Executive Officer,
American Psychological Association
Dianne C. Brown, Testing and Assessment Officer,
American Psychological Association

Council for Basic Education

A. GRAHAM DOWN
Executive Director

Congressman Augustus F. Hawkins
Chair, Committee on Education and Labor
B-346C Rayburn House Office Building
Washington DC 20515

20 June 1990

Dear Congressman Hawkins:

Thank you for your letter dated 6 June 1990 asking for views and advice on educational testing and assessment. This letter will summarize a ballooning literature on the subject. Supporting materials are available if you or your staff need them.

My own perspective is the effect of testing and assessment on curriculum and instruction--what children learn. The urgency for changing from multiple-choice, machine-scorable tests to performance assessments is fueled by the fact that the multiple-chocie tests trivialize the curriculum, reducing it to a series of unrelated facts which require children only to recognize them, not to use them. American society needs thoughtful adults who can solve problems, adapt to change, and use intelligently the resources of a technological world.

Chapter 1 students urgently need curriculum and instruction which is not driven by multiple-choice tests. It is a national tragedy that the Chapter 1 legislation mandates (and Department of Education regulations reinforce) a nationally normed and aggregatable test which at the moment must be multiple-choice. Many states are seeking alternatives, among them your own state of California. The California Assessment Program is proposing to use "grantback" money to design performance assessments for Chapter 1 which will both report accurately on the program and remove the obstacles to a thinking curriculum for Chapter 1 children. These efforts need encouragement from your committee.

Brief answers to the questions at the bottom of page 1 of your letter:

- Learning is best assessed by asking students to do what we want them to do--write, solve problems, display understanding. The means include: direct writing assignments scored by holistic scoring groups; portfolios; open-ended questions in mathematics and science; experiments and manipulations of equipment and materials; simulations, debates, and mock trials; problem-solving contests (the

725 Fifteenth Street, N.W. Washington, D.C. 20005 (202) 347-4171

Odyssey of the Mind, e.g.).

* Assessment programs at the Federal, state, and local levels could be integrated by a series of interlocking group grading sessions. Let me explain by building on the example of the California Assessment Program's highly successful grade 8 and grade 12 writing assessments. Each year, groups of teachers from across the state score the essays, which can be written on up to eight different topics, assessing ability to write in different real-life genres.

  Now imagine that other states in the Western Region have a similar writing assessment. Ten percent of the papers from each state are scored again (anonymously of course) by a group drawn from the states represented. The same process would go on in other regions of the country--Southeast, Atlantic States, Central, etc.

  Finally at the national level, 10 percent of the regional papers would be scored by a national committee.

  Why do this instead of expanding NAEP? For these reasons:

  - This is not an additional assessment--it uses existing state (and/or local assessments);
  - It involves teachers, administrators, parents in scoring groups, thus informing them directly about what students can do and should be able to do;
  - It is a bottom-up, not top-down, process, giving the people closest to the classroom ownership of a professional responsibility;
  - Because of the large number of people involved, information about standards is widely disseminated. How many people can cite the results of NAEP assessments now?

  The process is sometimes called "group moderation." It was proposed as a feature of the new English national assessment, but was not adopted or funded by the English government. The U.S. Congress has an opportunity to demonstrate educational leadership here.

* The adverse effects of testing can be minimized by phasing out multiple-choice, machine-scorable tests designed by test publishers.

  There are minimal adverse effects of performance assessments, since many assessments are no different from and in some cases better than ordinary classroom activities. The New York State Grade 4 students who took the science manipulative skills test in May 1989 and May 1990 wrote "Thank-you" on their papers and asked could they do the test again tomorrow.

- Comprehensive systems of assessment can be developed by expanding the pool of performance assessments and using psychometric expertise to develop sound scoring methods. Assessments should concentrate on program and school assessment, which means that matrix sampling can be used widely--not every student needs to be assessed. (However, some performance assessments like the New York State science tests are so intriguing that no-one wants to be left out!)

  Student assessment should not be used for selection and sorting. It should be developed as a profile of the student's strengths and weaknesses, with multiple indicators, never a single score.

- The Federal role in improving testing and assessment should be leadership, not regulation or imposition of top-down assessments like NAEP. The Federal government should specify educational outcomes and then assist states and localities to meet them.

  The Department of Education should be a resource, developing, researching, refining performance assessments. It should encourage experimentation at all levels and offer expert assistance to state and local education authorities seeking to make curriculum and assessment complementary. It has an obvious role in coordinating a national "group moderation," as described above.

The issue of cheating on tests (the focus of Cannell's books) is not relevant when tests are changed and become performance assessments. It is a red herring which distracts from the real issue. Cheating on tests has little to do with what children learn; it seems to be focused on exposing an irrelevant crime. The issues are teaching and learning, and ensuring that reasonable demands for accountability do not intrude on them or distort them.

I am available for further information and discussion of this and other educational issues.

Sincerely

Ruth Mitchell

Ruth Mitchell
Associate Director

AFT AMERICAN FEDERATION OF TEACHERS AFL CIO

ALBERT SHANKER
President

555 New Jersey Avenue, NW
Washington, DC 20001
202/879-4400

June 21, 1990

Honorable Augustus F. Hawkins
Chairman
Committee on Education and Labor
U.S. House of Representatives
Washington, DC 20515

Dear Mr. Chairman:

Thank you for your letter of June 6, 1990, requesting my views on the subject of educational testing and assessment. I would have been pleased to appear at a hearing on this subject because I agree with you that testing and assessment must be addressed if we are to improve our nation's educational performance. I hope you schedule additional hearings on this subject in the near future so that I can have the opportunity to explore this complex and important issue with you in greater depth than a written statement allows.

I am gratified that the Congress is taking an interest in the role and effects of traditional standardized testing on the quality of teaching and student learning in our nation's schools. Testing is a major enterprise in our education system, driving federal, state and local education dollars as well as instructional decisions. The nature and quality of the tests we use, and how we use them, are therefore of vital significance.

The 750,000-member AFT has long supported testing, chiefly for these reasons: We have no other comparably reliable means for determining if and how well the nation's youth is being educated and the extent to which our schools are discharging their public responsibility. In particular, we have no other means for measuring progress toward overcoming our past legacy of denying equal educational opportunity to poor and minority youngsters and for assessing the inequities that continue to exist. Moreover, the public deserves -- indeed, has a right -- to know what we are getting for our education dollars.

But while the AFT supports testing, we are critical of the quality of the tests most commonly used in our school system and the ways in which they are employed. Briefly summarized, the AFT, along with a growing number of testing

and education experts, has become convinced that these tests tend to narrow teaching and learning -- indeed, may have contributed to the "dumbing down" of America. Additionally, existing tests severely constrain promising education reform initiatives.

These problems associated with standardized testing deserve serious national attention and a commitment to developing reliable, publicly useful assessments that help promote educational achievement. Unfortunately, encouraging local districts to develop new assessments is not the best means to achieve that end. In fact, we fear that this well-intended measure would add another layer of testing and assessment to already overburdened students and teachers. We also do not believe that new, district-developed tests -- each of which would be different -- can yield trend data or comparable information, thereby exacerbating the existing problems in education reporting. Moreover, since the capacity of local school districts in alternative assessment is very thin, this measure is likely to add to the already plentiful supply of education hucksters that districts are prey to while reducing the impact of responsible groups, including some states, presently working on new assessments. Quality control in developing new assessments is, in short, essential.

Congress should also be aware that the U.S. Department of Education's Office of Educational Research and Improvement is in the midst of competing many of the federal research and development centers, a center on testing among them. Any new legislation affecting assessment ought to proceed in light of the results of that competition. It also would be appropriate for Congress to consider any assessment initiative in light of the national education goals adopted by the President and the Governors.

The AFT has offered responsible criticisms of the present testing system. We are eager to cooperate with legislative and other means to develop assessment systems that not only overcome the problems of the present systems but also help to stimulate needed improvements in educational achievement.

Nevertheless, the AFT urges caution when it comes to a local, district-based strategy for developing new assessments, especially without getting a handle on existing testing. We need to address the issue of standardized

testing as a nation with much at stake in the issue. Proposed solutions that diffuse authority and responsibility for developing a valid assessment system could have tragic consequences for our educational system.

I look forward to further dialogue with you on this critical issue.

Sincerely,

Albert Shanker
President

AS/dr
opeiu2aflcio

The Need for a National Assessment of Educational Progress in Foreign Language Competence

by
Daniele Ghiolfi Rodamar
Assistant Professor
American University

Oversight Hearing on
Testing in Education
The Subcommittee on Elementary,
Secondary, and Vocational Education

2175 Rayburn H.O.B.
U.S. Congress
June 7, 1990

Department of Language and Foreign Studies
4400 Massachusetts Avenue, N.W., Washington, D.C. 20016-8045 (202) 885-2381

Mr. Chairman, members of the Committee, this morning I am honored to present testimony. My name is Daniele Rodamar. I am an Assistant Professor of French literature and language at American University in Washington, D.C. The following testimony reflects my experience as a university level foreign language instuctor for over a decade and as a faculty member with responsiblility for foreign langauage curriculum development, program coordination, and assessment for elementary and intermediate French language courses. I am speaking as an individual, and my testimony does not necessarily represent the views of America University.

DEMAND FOR FOREIGN LANGUAGE SKILLS IS GROWING

Today's kindergarteners will graduate to a world that will provide many opportunities to put foreign language skills to work. Language education is a fundamental element of curricula in our nation's schools. As Bill Honig, California's Superintendent of Public Instruction said in launching a campaign to strengthen California's K-12 language education: "Learning a foreign language opens many doors for students. It allows them to compete in an international job market where proficiency in another language is no longer a luxery but a necessity. They also better understand our own diverse society and develop communication skills necessary to expand their perspectives of the world." The trends in trade, foreign investment, international tourism, the increasingly global organization of business and other factors are increasing the need for foreign language skills.

INFORMATION ON ACHIEVEMENT IS MISSING

How are we doing in strengthening America's foreign language

education? While there is some data on "process" variables, such as enrollments, "seat time", the number of foreign language teachers and so on, we know little at the national level about the proficiency of the students who graduate from these language programs. Anecdotes (such as the efforts to sell the "no go" NOVA Chevy in Latin America, the Pepsi "Come Alive" ad campaign that failed in Thailand when it was translated as, "It brings your ancestors back from the dead", and President Carter's speech that told of his "lust" for the Polish people) suggest that all is not well. Two thirds of the translating jobs at the U.S. Department of State are filled by foreign-born individuals because properly trained American-born candidates are not available. The pattern in the private sector does not appear much better. The snapshots of language proficiency provided by various studies reinforce these concerns about the foreign language proficiency of America's students.

Assessment of educational progress is a fundamental element in strengthening educational achievement. This has been recognized by the Coalition for the Advancement of Foreign Languages and International Studies (CAFLIS) which represents 165 member organizations from all levels of education, the business community, state and local governments, language and exchange groups, and others. CAFLIS has called for assessments of progress in foreign languages and international studies as part of a plan of action for upgrading foreign language and international studies education. Assessments of foreign language achievement should be a mandated element of the National Assessment of Educational Progress (NAEP). If done in a responsible and methodologically sound manner, a national assessment of educational progress in foreign languages will encourage improvements in language education not only by providing information on how we are doing but also by spotlighting the importance of foreign language education to the

nation and sending a clear signal that foreign language needs to be a core element in the curricula of our nation's schools in the elementary and secondary level as well as in our nation's colleges and universities.

A NATIONAL OBJECTIVE: A LANGUAGE COMPETENT AMERICA:

There has been growing awareness of the need to strengthen foreign language education in the United States. In November 1979 the President's Commission on Foreign Language and International Stuides pointed with alarm to our citizens' lack of international knowledge. As the Chairman of the Commission noted in transmitting the study to President Carter, "the hard and brutal fact is that our programs and institutions for education and training for foreign language and international understanding are both currently inadequate and actually falling further behind. This growing deficiency must be corrected if we are to secure our national objectives as we enter the Twenty First Century." By the mid-1980s reports calling to strengthen foreign language eduction began to be made by groups with the power to actually influence events in our schools, such as the Council of Chief State School Officers, the National Governor's Association, and the Southern Governors' Association.

Earlier this year, following the Charlottesville "Education Summit", the President and the nation's governors agreed to six major goals and twenty six objectives for educational improvement by the year 2000. Two objectives relate directly to second language study; others are more indirectly related. The President and the Governors gave high priority to the development of quality assessments to monitor progress toward these educational goals and objectives.

In brief, the increasing importance of foreign languages for U.S. security, prosperity, and growth has been increasingly recognized by leaders in education, business and government.

TRANSLATING OBJECTIVES INTO ACTION:

The increased emphasis on language skills has been accompanied by growing enrollments. Recent surveys conducted by the Joint National Committee for Languages found that 30 states have instituted or increased foreign language requirements in the last ten years. Recent figures indicate a ten percent increase in foreign language enrollments during that same period.

The impact of these changes is just beginning to be felt. For example, the state of New York's Action Plan to Improve Elementary and Secondary Education Results includes a commitment to second language education for all students. Beginning with the class of 1994, all students will take at least two years of a second language prior to grade 9 and additional incentives for continuing language study are made in the form of requirements for the Regents' Diploma. In California--which has as many students as the smallest 24 states combined--the Hughes-Hart Education Reform Act of 1983 mandated one year of foreign language study as an option to meet high school requirements. California's public universities have required at least two years of study of a single foreign language for admission, and the state's Board of Education has recommended that all high school students complete two years of study in a foreign language. California's enrollments in foreign languages grew by a third between 1981 and 1987--but only 14% of the students in kindergarten through 12th grade were enrolled. In brief, important changes have been initiated and their full impact will be felt in coming years. National level information on trends in achievement in second language proficiency that can be disaggregated to at least the state level is vital in building effective foreign language programs.

THE NEED FOR NATIONAL ASSESSMENT IN FOREIGN LANGUAGES

University teachers already have high school transcripts and advanced placement tests to know know what language skills

students are bringing to campus. In a few states, state assessments of foreign language achievement add information.

While this may be enough to create a wall-chart, this leaves postsecondary faculty, as well as K-12 faculty without a clear picture of how the system as a whole is working. In some subject areas there is not even a yardstick of achievement: the College Board provides widely used achievement tests in German, French, Spanish, Italian, Hebrew, and Latin--but not in any of the Asian languages. This forces each postsecondary institution to provide its own hit-or-miss assessment and sends out a signal to students, parents, teachers, and administrators about the relative importance of languages.

A national assessment of educational progress in foreign languages is important in getting authorizations and achieving funding for foreign language education. The monitoring of achievement by a National Assessment of Educational Progress in Foreign Languages would spotlight "how we are doing" and would send a clear signal that results matter.

We face major problems in our efforts to improve language education. Too often teachers who have very limited proficiency in the language they are supposed to be teaching operate without effective training, feedback and support. The need to fill elementary and secondary classrooms with "a warm body" often pre-empts questions about the results. This absence of quality information on how we are doing makes it difficult to drive improved program performance and improved articulation across grade levels. The picture is further clouded when teachers pressed with the need to keep students, parents, and administrators happy allow grades to creep upward without corresponding increases in achievement. The problem is not grade inflation by individual teachers. It is more serious than that. We simply do not know how the system is working and this lack of information moves the emphasis to process rather than results.

Without information on how the whole system is working, there is little systematic pressure to upgrade the quality of teaching, and to provide the funding needed for materials, salaries, and articulation across grade levels.

A national assessment of educational progress in foreign languages, would provide information on how the nation as a whole is doing--and on how one state or region is doing relative to another. Such information can play a key role in driving dissemination of effective programs, building support for adequate funding and improving articulation across grade levels. The requisite consensus building process can help ensure that programs reflect the language competency needs in business, industry, agriculture, the professions and government, as well as in teaching and research.

Today we have too little language education too late in the educational program. Information on foreign language acheivement of students at lower grades would provide a fulcrum for leveraging improvement and for providing a more realistic time table for students to learn foreign languages. This is no small matter. As California's Foreign Language Framework put it, "No matter how good the pedagogy, students will not become fluent in a second language by attending a 50 minute class five times a week during a single school year. Mastery of foreign language takes time. (In Europe, Japan, and the Soviet Union, for example, five to seven years are generally allocated to the study of English or another foreign language.) For school administrators interested in building a successful language program, the requirement for a large block of time has two clear implications: First, it signals the need to move the beginning of the serious study of language into the kindergarten thrugh grade eight years. And second, it highlights the importance of district wide strategic planning so that continuity of learning is not interrupted." A national foreign language assessment

would help draw attention to these issues.

For university and college teachers, such as myself, this information would provide a useful basis to work in academic alliances to upgrade K-12 language education. The process of assessment and interpretation would force K-12 and postsecondary education bureaucracies to face the issue of what they are doing and what the results are. In teacher training it would help provide vital systematic feedback on how the people we are turning out with degrees are doing when they find themselves in front of a classroom full of typical American kids. This is system level feedback that teacher certification or other process variables cannot provide.

In sum, a national assessment of educational progress in foreign language education provides information on how the system is doing and sends out a signal that language matters and is a vital part of the curriculum. State by state and other comparisons properly conducted can aid in identifying and disseminating models of effective language education. And information on what other students are achieving can provide useful information to students that motivates their language learning strategies.

THE FOUNDATIONS FOR LANGUAGE ASSESSMENT ARE IN PLACE

A substantial portion of the research and development necessary to institute a national assessment of educational progress in foreign languages is already underway. While foreign language education--like other areas assessed by NAEP--seeks to build a complex of skills and to achieve a variety of goals, guidelines for the assessment of foreign language proficiency have been developed by the American Council of Teachers of Foreign Languages (ACTFL). Pressed with the need for assessment of foreign language proficiency, the U.S. government has long conducted assessments of language competency for use in placement and as a guide to future training. The state of Connecticut has

already conducted its own assessment of foreign language proficiency in its schools. The Educational Testing Service has long provided foreign language achievement tests for use in placement of students entering postsecondary education.

There is already action to move beyond this. The American Council on the Teaching of Foreign Languages and the Educational Testing Service, working with the testing descriptions developed by the U.S. Government Interagency Language Roundtable, have initiated efforts to forge a consensus among language educators regarding proficiency standards appropriate to traditional settings.

NAEP is the appropriate location for an assessment of foreign language achievement. For over two decades NAEP has provided valuable information at the national level on the quality of educational achivement. The often troubling results of these assessments--along with other streams of information such as ACT and SAT scores, dropout rates, reports from employers, and so on--have helped trigger and sustain the school reform movement. NAEP is the only regular national level assessment of achievement in core curriculum areas. Under the Hawkins-Stafford Act (PL 100-297) NAEP has been expanded to provide a wider range of comparisons across core curricula. Adding foreign languages to the assessments of NAEP would build on an established institution and would send a powerful signal regarding the centrality and importance of achievement in foreign language education.

CAVEATS:

America has benefited from having a highly decentralized system of education which allows for diversity in goals and approaches and encourages flexiblity in meeting local needs. Many Americans have viewed national level assessments of achievement with extreme caution, aware that no assessment can

measure everything--and that what is left out may be as important as what is included. While an assessment may be more or less "curriculum neutral", all assessment scores have a necessary correlation with curriculum. In a field as diverse as language education, this does not reduce the need and value of assessment but it warns against over interpretation of results. The diversity is quite real. A 1976 study by the Articulation Council Liaison Committee on Foreign Languages found that not even an area perceived to be as central to language instruction as vocabulary was standardized. Among 28 elementary and intermediate German texts examined, less than five percent of the total words listed were common to all texts. Subsequent studies showed that student and K-12 teacher perceptions of what was expected in postsecondary programs varied greatly. The dynamics of consensus about what is important in the rapidly changing field of language education makes it strongly advisable not to attach too much weight to any single measure.

The applicability of a national level assessment for judging the success or failure of individual state or local level program reforms is at best questionable because assessment scores may change for reasons havin little or nothing to do with the assessment, including changes in student backgrounds and curriculum alignment. This is another reason why NAEP complements other information (such as state assessments and SAT scores and postsecondary education or employment outcomes) on how we are doing. If truth is, as one methodologist claimed, the convergence of independent streams of data, then it is vital that in a dynamic and diverse system such as our own that this diversity of approaches and measures be preserved. While it is appropriate that NAEP inform education debates and programs, the necessary imperfections of measurement by any single instrument and the importance of encouraging constructive debate among researchers, teachers, parents, and students make it essential that NAEP continue to complement other data streams rather than preempting

or defunding them.

NAEP is valuable in providing an assessment that no one aligns curriculum to meet. It informs rather than coerces, and as such fits with the best traditions of our nation's education system. It is important that NAEP continue to be used in ways that inform education, that strengthens rather than undermines education. Language education is multidimensional and pursues mulitiple goals: NAEP must acknowledge this. Multiple choice tests are helpful, but not enough. NAEP must continue to move toward improved and authentic assessment. Assessing competency in a language is not the same as testing achievement. Achievement tests are constructed to check mastery of some discrete body of material covered in a course of instruction. They provide feedback, but they typically test for specific, often unconnected elements of language. A competency test on the other hand is a holistic assessment of what the student can actually do with the language in a unrehearsed situation. The student's response to a testing prompt is not simply right or wrong; it is indicative of a stage of competency and helps define the student's performance level. A competency test addresses what can be done now. NAEP has emphasized these issues of competency in other assessment areas, and should do so in the area of language as well. Process and context cannot be ignored if we want to know how programs are working. That is why information on variables such as access to language education technology and proficiency of Hispanic students studying Spanish, and teacher proficiency should be provided.

In brief, while NAEP should be part of a larger system of research and feedback, it can provide a useful contribution that will play a critical role in improving language education in our nation.

What matters to the nation in language assessment is the level of proficiency of students in using a second language.

Since the ability to use language skills in real world contexts is the priority, the assessment of foreign language skills should focus on foreign language proficiency: on the ability to use language rather than their achievement's in reciting the vocabulary or syntax of any particular textbook or group of textbooks. The increasing emphasis in many classrooms on real world interactions--through telecommunications, non-textbook printed materials and so on make this emphasis on 'proficiency', rather than on 'achievement' in the narrow sense, particularly important. This is consistent with the approach used in assessments, such as the NAEP reading assessment, which emphasizes assessment of the reading skills needed to function in today's America.

The variety of languages studied across our nation and the costs of assessment pose the difficult issue of which languages to assess. The large majority of American students study Spanish and French. Other languages, such as Arabic, Chinese, Japanese, and Russian are studied by relatively few students but may be deemed of national interest for strategic, economic, and other reasons. Here again the consensus building effort that characterizes NAEP are particularly appropriate for determining which languages to assess and with what periodicity.

CONCLUSION:

The establishment of a national assessment of educational progress in foreign languages would be important because: (1) it would provide vital information to students, teachers, and others about how foreign language education programs are working, (2) it can help identify foreign language programs that work and strengthen the ability to disseminate those programs, and (3) it sends out a clear signal to students, parents, teachers, administrators, legislators and others that language education at the K-12 level is an essential part of the curriculum and that

competency matters.

National assessments of achievement in foreign languages are an essential tool in upgrading the quality of foreign language instruction. I respectfully urge the members of this committee and of the U.S. Congress to mandate foreign language assessment as a regular component of NAEP. This--along with ongoing input from the field and regular Congressional oversight--can play a vital role in upgrading language instruction, in helping to meet the national goals in education, and in allowing America to transform the many challenges that face us in this rapidly evolving world economy into opportuntities.

Thank you.

STATEMENT OF

FREDERICK H. DIETRICH

VICE PRESIDENT FOR GUIDANCE, ACCESS, AND
ASSESSMENT SERVICES

THE COLLEGE BOARD

TO

COMMITTEE ON EDUCATION AND LABOR
U.S. HOUSE OF REPRESENTATIVES

WASHINGTON, D.C.

JUNE 29, 1990

Members of the Committee on Education and Labor, I am Fred Dietrich, Vice President for Guidance, Access, and Assessment Services at the College Board. I very much appreciate the opportunity to comment upon testing, assessment and evaluation issues currently being considered by the Committee on Education and Labor.

Founded in 1900, the College Board is a national nonprofit association of more than 2700 colleges and universities, secondary schools, school systems and education associations and agencies. The Board assists students who are making the transition from high school to college through services that include guidance, admissions, placement, credit by examination and financial aid. In addition, the Board also sponsors research, provides forums to discuss common problems of education and addresses questions of educational standards.

The College Board firmly believes that quality assessment of student skills and achievement is ultimately crucial to the long-term social, economic, and political well-being of the United States. Nothing is more important to our future economic growth and social progress than education of the highest quality. Used sensitively, instruments of assessment can help achieve that end.

Over the last decade, the issue of standards and expectations of students has been a particular focus of the College Board's Educational EQuality (EQ) Project. EQ's efforts in the first part of the 1980s resulted in a set of publications describing "what students should know and be able to do" on graduating from high school. Academic Preparation for College, known as the Green Book, describes learning outcomes for high school curricula in six basic academic subjects--English, the arts, mathematics, science, social studies, and foreign language. It also identifies basic academic competencies--reading, writing, speaking and listening, mathematics, reasoning, and studying--which depend on, and are further developed by, work in these subjects. The "rainbow" series goes further in providing specific curriculum and instructional suggestions about how to achieve the results outlined in the Green Book.

EQ's work has involved consensus building among teachers. Hundreds of educators from both schools and colleges helped to compile the Academic Preparation series. This series does not address specific grade levels but rather the learning outcomes which should result from a student's exposure to a full educational experience through twelve grades.

The consensus of educators involved with EQ is that much of the Green Book, and in particular the basic academic competencies, are appropriate for both college- and work-bound students. We believe it is important to promote high academic standards for all students, rather than setting minimum competencies for most and tougher expectations for some. The goal should be to give all high school students access to the knowledge and skills necessary for entering and completing higher education. Some may not go to college right away, but we should try to keep their options open. Moreover, employers have told us that the EQ basic competencies are what they need in new hires. In terms of basic skills, there may be little difference between what is needed by the college-bound and those headed for employment.

Also particularly relevant to your deliberation are two brochures--The Educational Equality Project and College Board Examinations and Improving Academic Preparation for College: The Role of Assessment--which address the "congruence" between our tests and the EQ-defined competencies and skills. We will be pleased to provide all these items to the Committee.

Several College Board instruments could be helpful to your present discussion:

- Descriptive Tests of Language and Mathematics Skills--designed to assess the battery of skills (writing, thinking, reading, analysis, and mathematics) that students must have to perform well at the college level, closely aligned with the goals described in Academic Preparation for College.

- The Advanced Placement (AP) Examinations--a program of college-level courses and examinations for secondary school students in 16 disciplines. About 37 percent of American secondary schools currently participate in the program, serving approximately 17 percent of their college-bound students. Those of you who have seen the movie "Stand and Deliver" will know how important and valuable this program can be for minority students.

- The Achievement Tests--a series of 15 tests in 14 subject areas taken by some 300,000 college-bound students each year and designed to measure knowledge, and the ability to apply that knowledge, in specific subject areas.

- The Scholastic Aptitude Test (SAT)--a nationally administered test that measures developed verbal and mathematical reasoning abilities related to successful performance in college, taken by some 1.8 million students each year.

- The Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT)--a school-based test that measures verbal and mathematical reasoning abilities important for success in college, taken by more than 1.5 million high school sophomores and juniors each year.

You may also be interested to know that the College Board through its office in Puerto Rico sponsors the Prueba de Aptitud Academica (PAA), sometimes referred to as the "Spanish SAT." The PAA is taken by over 100,000 students throughout Puerto Rico, Latin America, and the mainland United States. Not a translation of the SAT, the PAA is composed of items developed directly in Spanish; like the SAT, the PAA measures two essential types of reasoning: verbal and mathematical. Along with the PAA, we also administer a battery of subject-matter achievement exams in Spanish.

These and other standardized tests can be very useful in evaluating what students have learned. When properly developed, using the knowledge of teachers and other curriculum experts as well as surveys of appropriate curriculum and course content, standardized tests can measure many of the important learning objectives that schools have for themselves and for their students, and do so validly, efficiently, and inexpensively.

Finally, I should note modes of testing other than traditional paper-and-pencil multiple-choice tests. Clearly not all knowledge can be measured by these traditional tests. You may be interested to know that we are currently exploring a number of modifications to the SAT that include the addition of open-response items in which students solve a problem and record their answers directly (that is, not via multiple-choice items), as well as a writing component (and score) that could include an essay or other direct measure of writing ability. These explorations also include the development of what we call "proficiency scaling," through which additional information will be generated about what particular scores on the SAT (in its verbal, mathematics, and writing components) mean in terms of what students are able to do.

Perhaps the most promising news of all in efforts to measure what individual students know and are prepared to do is the development of computer-delivered tests. The College Board's first application of computerized adaptive testing has been in a series of tests of skills in college English and mathematics known as Computerized Placement Tests. The program is being expanded to include assessment of mathematics at higher skill levels. We are also investigating other applications of computerized adaptive testing, including a battery of assessment tools and accompanying guidance materials for use with students at the middle school level and those with limited English proficiency.

What is so encouraging about this kind of test is that it can utilize student responses to previous questions to select later questions in order to more accurately describe individual student's abilities and needs. It's almost a different test for each student, created by the student's own level of ability and knowledge. These tests will require much less time to take than paper and pencil tests, and will provide the option of immediate scoring and feedback to facilitate counseling, course placement, and other forms of advisement, and provide more useful diagnostic information.

The College Board is pleased to offer assistance in using existing tests and/or in developing additional ones. As I have tried to describe in this statement, the College Board has long experience in measuring higher order thinking skills, in using tests to inspire advanced levels of learning, and in setting common educational standards and goals through consensus-building activities.

Thank you again for the opportunity to present this statement. We look forward to working with the Committee on Education and Labor on these important educational issues.

6439j/6014j